ABSTRACT

Title of dissertation: RESILIENCY ASSESSMENT AND ENHANCEMENT OF INTRINSIC FINGERPRINTING

Wei-Hong Chuang, Doctor of Philosophy, 2012

Dissertation directed by: Professor Min Wu
Department of Electrical and Computer Engineering

Intrinsic fingerprinting is a class of digital forensic technology that can detect traces left in digital multimedia data in order to reveal data processing history and determine data integrity. Many existing intrinsic fingerprinting schemes have implicitly assumed favorable operating conditions whose validity may become uncertain in reality. In order to establish intrinsic fingerprinting as a credible approach to digital multimedia authentication, it is important to understand and enhance its resiliency under unfavorable scenarios.

This dissertation addresses various resiliency aspects that can appear in a broad range of intrinsic fingerprints. The first aspect concerns intrinsic fingerprints that are designed to identify a particular component in the processing chain. Such fingerprints are potentially subject to changes due to input content variations and/or post-processing, and it is desirable to ensure their identifiability in such situations. Taking an image-based intrinsic fingerprinting technique for source camera model identification as a representative example, our investigations reveal that the finger-

prints have a substantial dependency on image content. Such dependency limits the achievable identification accuracy, which is penalized by a mismatch between training and testing image content. To mitigate such a mismatch, we propose schemes to incorporate image content into training image selection and significantly improve the identification performance. We also consider the effect of post-processing against intrinsic fingerprinting, and study source camera identification based on imaging noise extracted from low-bit-rate compressed videos. While such compression reduces the fingerprint quality, we exploit different compression levels within the same video to achieve more efficient and accurate identification.

The second aspect of resiliency addresses anti-forensics, namely, adversarial actions that intentionally manipulate intrinsic fingerprints. We investigate the cost-effectiveness of anti-forensic operations that counteract color interpolation identification. Our analysis pinpoints the inherent vulnerabilities of color interpolation identification, and motivates countermeasures and refined anti-forensic strategies. We also study the anti-forensics of an emerging space-time localization technique for digital recordings based on electrical network frequency analysis. Detection schemes against anti-forensic operations are devised under a mathematical framework. For both problems, game-theoretic approaches are employed to characterize the interplay between forensic analysts and adversaries and to derive optimal strategies.

The third aspect regards the resilient and robust representation of intrinsic fingerprints for multiple forensic identification tasks. We propose to use the empirical frequency response as a generic type of intrinsic fingerprint that can facilitate the identification of various linear and shift-invariant (LSI) and non-LSI operations.

RESILIENCY ASSESSMENT AND ENHANCEMENT OF INTRINSIC FINGERPRINTING

by

Wei-Hong Chuang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Min Wu, Chair/Advisor
Professor K. J. Ray Liu
Professor Rama Chellappa
Professor Gang Qu
Professor David Jacobs

To my family.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Min Wu, for her guidance and support over the past five years. She taught me using numerous examples how to identify important problems, how to conduct practical engineering research with solid insights, how to think creatively while being rigorous, and how to set up the proper mindset when tackling challenges in research. Her constructive suggestions and comments have substantially improved my writing and presentation skills, and her wise ways of dealing with real-world issues have been invaluable lessons to me. Perhaps most importantly, her pursuit of excellence in research and in all aspects of life has meant so much to me and inspired me to raise my standards, too. I will keep in mind all the principles and experiences that I learned from her in my future endeavors.

I also want to thank Prof. K. J. Ray Liu and Prof. Rama Chellappa for their wonderful courses and their unwavering help during my graduate study. Their knowledge and vision have been great resources to me. Likewise, I thank Prof. Steve Marcus for showing me what an educator is about. I also want to thank Prof. Gang Qu and Prof. David Jacobs for serving on my dissertation committee and for their valuable comments. I also appreciate the mentorship and financial support offered by the Future Faculty Program at Maryland.

It is my privilege to have worked with outstanding colleagues at Maryland. I want to sincerely thank Dr. Ashwin Swaminathan and Dr. Avinash Varna for giving me a lot of help and guidance. My daily discussions with Dr. Varna and Dr. Wenjun Lu have been very interesting, encouraging and helpful. I also thank Dr. Lu's help during my job search. It is also a great pleasure to spend a lot of time with MAST and SIG members, in particular Ravi Garg, Hui Su, Chau-Wai Wong, Adi Hajj-Ahmad, Dr. Wan-Yi Lin, Dr. Yan Chen, Dr. Matthew Stamm, and Yu-Han Yang. Certainly, those good friends that I made at Maryland have been very important to me. Special thanks go to Wei-Hsuan Yu, Mei-Chun Lu, Yu-Jen Chen, Pinni Chung, Yi-Ting Hou, Tsung-Hsueh Lee, Hsiang-Hwang Wu, Ling Hung, Jen-Chien Chang, Dr. Chia-jung Tsui, Dr. Yen-Chen Liu, Yu-Ting Kao, Dr. Nai Ding, Dr. Xinxin Yang, Hami Siahpoosh, and Akiko Nakayama for their care and for all the good times we had together. Friendship never vanishes.

Last but definitely not least, I thank those that love me and I love. I give my sincerest gratitude to my parents who have been supporting every action throughout my life. They have been so comforting every time when I got frustrated. I also owe my deepest thanks to my wife Chia-Wei, without whose love, sacrifice, and companionship this dissertation would have been impossible. I feel extremely fortunate to share my life with her. I am also grateful to my sister, my brother, and every one in my extended family for their constant encouragement. I dedicate this dissertation to them.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Digital Multimedia Authentication using Intrinsic Fingerprinting

Recent advancement of multimedia and communications technologies has significantly facilitated the creation and distribution of digital multimedia data, such as images, videos, and music. Compared to multimedia signals in analog forms, digital multimedia data have the significant advantages of easy acquisition, storage, and transmission, and therefore have been widely used in various applications where multimedia content is involved.

Along with the growing importance of digital multimedia data, concerns regarding their misuse have also been raised and are receiving increasing attention. In particular, the digital nature of such data makes them easy targets for manipulations, and a large amount of data have been found tampered or forged so as to convey misleading or false messages [22]. In order to establish the credibility of digital multimedia data, it is crucial to devise mechanisms that can examine their integrity or further infer the processing history that they have gone through. Many solutions have been proposed in the recent literature toward ensuring that digital multimedia data are used in a trustworthy and authorized manner. In this dissertation, we focus on one particular class of strategies, commonly referred to as *intrinsic fingerprinting*. Intrinsic fingerprinting [21, 24, 52, 70] aims at exploiting certain "intrinsic traces" that have been left in the digital multimedia signal as it undergoes a processing pipeline. Such traces can be used to expose certain properties or patterns introduced by user manipulations, and thus are helpful in assessing the integrity of multimedia data.

1.2 How Resilient are Intrinsic Fingerprints?

Among various intrinsic fingerprinting techniques that have been designed and experimentally tested by the research community, many are based on statistical traces originated from certain signal characteristics that are subtle in nature. Fig. 1.1 illustrates a general setting of intrinsic fingerprinting. The source signal undergoes a chain of n_1 processing modules and reaches the point A. Intrinsic fingerprints created in this processing chain may be estimated by examining the features derived from the signal at A. In reality, however, extra post-processing may be performed after A, and only the final output at the point B is available for forensic feature extraction

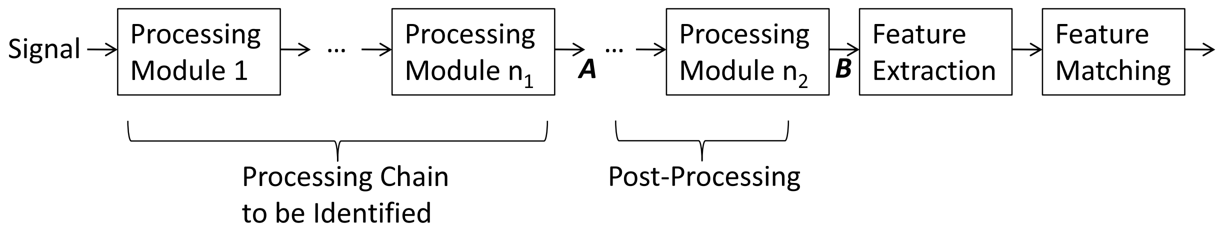


Figure 1.1: A typical setting of intrinsic fingerprint extraction and matching.

and matching. Even if the processing chain to be identified is kept unchanged, the extracted forensic features may depend on attributes of the source signal and are subject to changes if the post-processing causes the signals at points A and B to be different. It is therefore desirable that intrinsic fingerprints can be robustly identified against a variety of content characteristics and post-processing.

In addition, intrinsic fingerprints may also be extracted and matched in the presence of adversaries that are motivated to perform certain "anti-forensic" operations so as to counteract or mislead forensic analysis. Compared to the aforementioned content variations or post-processing, anti-forensics involves manipulations of the intrinsic fingerprints that are conducted by the adversaries. Therefore, resilient intrinsic fingerprinting against anti-forensics should take into account the interaction between forensic analysts and adversaries, and suitable countermeasures should be devised accordingly.

Finally, for many current intrinsic fingerprinting systems, the employed intrinsic fingerprints are tailored to particular forensic tasks such as JPEG compression or filtering. As such, multiple forensic features may need to be computed and matched in order to identify a processing module that is unknown to a forensic analyst. Such computation can be costly and reduce the practical usability of intrinsic fingerprints, and another aspect of resiliency concerns finding an intrinsic fingerprint that can identify a wide range of processing modules.

1.3 Main Contributions and Dissertation Organization

In order to improve the foundation and practical usability of intrinsic fingerprinting, this dissertation addresses several resiliency aspects of intrinsic fingerprinting. In particular, we examine current intrinsic fingerprinting schemes in terms of their resiliency to possible sources of fingerprint distortions, and propose solutions for resiliency enhancement.

1.3.1 Imaging Device Identification against Content Dependency and Post-Processing

Information about imaging mechanisms of digital images and videos carry useful clues about their origin, and therefore can serve as important intrinsic fingerprints for forensic analysis. However, most research so far has assumed that these fingerprints are extracted under favorable conditions, such as with controlled image/video content, native spatial resolution, and light to moderate compression. When such conditions are not met, it is still unclear how well the extracted fingerprints can be used to match the imaging mechanisms. In this dissertation, we investigate the resiliency of imaging device identification based on color interpolation and imaging noise traces against content dependency and post-processing. First, we show that the color interpolation coefficients that were proposed to represent the color interpolation algorithms have a substantial dependency on the image content. Such a dependency may cause mismatch between the coefficients estimated from training and testing data, and therefore reduce the identification accuracy. In order to mitigate the mismatch, we propose profiles that can be efficiently calculated from the image and can represent the image content, and then propose training image selection schemes based on these profiles for configuring a suitable classifier that employs training images whose content match the testing image so that the identification is significantly improved.

We also show that such post-processing as strong compression that can be found in low-bit-rate video applications has a considerable impact on the accuracy of camera unit identification using the Photo Response Non-Uniformity (PRNU) derived from imaging noise. As such, strong compression poses challenges to video-based camera unit identification as low-bit-rate videos become increasingly prevailing. Nevertheless, we have found that even within the same video, the compression level actually depends on the frame type, and by properly exploiting the difference in compression levels during the training and testing phases, we can achieve a substantially higher identification accuracy.

1.3.2 Anti-Forensics and Countermeasures of Color Interpolation Identification and Electrical Network Frequency Analysis

Anti-forensic operations intentionally manipulate the fingerprint extraction and matching, and can undermine the effectiveness of intrinsic fingerprinting. In this dissertation, we examine the resiliency against anti-forensics of two types of intrinsic fingerprinting: 1) color interpolation identification, which is a core technique used in camera model identification, and 2) electrical network frequency analysis for space-time localization of digital recordings.

We first investigate plausible anti-forensic operations that can be performed by such adversaries as pirate camera manufacturers. These operations include parameter perturbation that circumvents the identification of targeted interpolation algorithms, and algorithm mixing that can mislead the identification toward a specific wrong result. Our findings provide insights into the vulnerabilities of color interpolation identification based on gradient direction classification, and such insights motivate forensic analysts to take countermeasures, which may in turns be countered by adversaries' follow-up actions. We characterize such a dynamic interplay using game-theoretic techniques, and derive the optimal strategies that both sides are willing to adopt.

We then explore the anti-forensics of a recently developed class of space-time localization techniques based on the electrical network frequency (ENF). These techniques extract the ENF signal from a sensor recording and compare it to the references measured from the power mains to determine the creation time and region of the recording. While this technique has received increasing attention lately, its resiliency against anti-forensics has not been investigated. We establish a mathematical framework that can characterize plausible anti-forensic operations for ENF signal manipulations. This framework also motivates countermeasures against anti-forensic operations. We further consider possible improvements over anti-forensics that may evade the detection, which consequently call for refined forensic detection schemes. Such an interplay between forensic analysts and adversaries can be viewed from an evolutionary perspective and a game-theoretic perspective, and we study representative cases to obtain a quantitative understanding of such an interplay and to obtain optimal forensic analysis strategies.

1.3.3 Empirical Frequency Responses as Generic Intrinsic Fingerprints

In addition to determine the creation mechanism and time of digital multimedia data, another important goal of intrinsic fingerprinting is to discover the processing history that the multimedia data has undergone. As discussed earlier, current intrinsic fingerprints are often tailored to recognizing particular processing modules, and often fail when applied to other modules. Even if multiple fingerprints can be extracted and matched, they may still be unable to accommodate unforeseen operations. This leads to significant computational burden and limited effectiveness for forensic analysis.

We propose in this dissertation to use the empirical frequency response (EFR) as a generic intrinsic fingerprint. We show that many classes of image processing operations, either linear and shift-invariant (LSI) or non-LSI, such as resampling, JPEG compression, and non-linear filtering, exhibit distinctive patterns in their EFRs and therefore can be identified using the EFR representation. The EFR can also be used for other use in forensics. For example, we have found that EFR has some dependency on the model of the camera that is used to generate an image, and such dependency can facilitate camera model identification.

1.3.4 Dissertation Organization

The rest of the dissertation is organized as follows. In Chapter 2, we investigate the content dependency of camera model identification based on color interpolation identification. To mitigate the penalty incurred by the mismatch between training and testing images, we propose profiles that can be used to represent the image content type, and training image selection schemes that can automatically determine the training images that match the testing image.

In Chapter 3, we study another aspect of color interpolation identification, namely, its resiliency to anti-forensic operations including parameter perturbation and algorithm mixing. Our analysis sheds light on the inherent vulnerabilities of current color interpolation identification schemes. We propose a color interpolation identification game to characterize such an interplay between forensic analysts and adversaries.

In Chapter 4, we continue to investigate the impact of anti-forensics when applied to a recent time-stamping technique based on the electrical network frequency (ENF). We show that certain anti-forensic operations can manipulate the ENF signal, which can be detected under our mathematical framework by examining appropriate types of consistency. Improvements by adversaries as well as refined forensic techniques can arise from an evolutionary perspective. We characterize such a dynamic interaction using game-theoretic techniques, and quantitatively evaluate representative scenarios and determine the optimal strategies.

In Chapter 5, we study the resiliency of intrinsic fingerprints against post-processing and present a case study of imaging-noise-based camera identification using strongly compressed videos. As such compression reduces the identification performance, we show that within the same video there exists different levels of compression, which can be leveraged to improve the identification using a fixed number of video frames.

In Chapter 6, we consider the applicability of intrinsic fingerprints to a wide range of forensic tasks, and propose to use the empirical frequency response (EFR) as a generic intrinsic fingerprint. We show that the EFR can identify processing modules that are either linear and shift-invariant (LSI) or non-LSI and can facilitate the identification of camera models.

Finally, we conclude in Chapter 7 this dissertation and outline research issues that can be explored in the future.

CHAPTER 2

Content Awareness for Camera Model Identification

2.1 Chapter Introduction

In the past decade, the rapid advancement of digital photography, storage, and Internet technologies has boosted the ubiquitous use of digital images in today's society. In the meantime, since digital images are vulnerable to software editing and manipulations, increasing attention has also been brought to concerns regarding their origin and integrity. One can readily ask a series of questions about a given digital image: for example, what kind of acquisition device was used to generate this image? If the image was taken by a camera, what is the make and model of the camera? Has this image undergone any non-trivial post-processing or manipulation?

All these questions lie under the umbrella of *digital image forensics*, which has become a very active research area in recent years. Extensive efforts have led to a number of promising techniques and tools. Fridrich *et al.* [24] developed the methodology of exploiting the Photo-Response Non-Uniformity (PRNU) to distinguish different camera units. Swaminathan *et al.* [67] showed how to employ color interpolation coefficients robustly and use them to identify different camera models. Also, they employed blind deconvolution to estimate a linear and shift-invariant (LSI) approximation of the overall post-processing step and the LSI estimate will be matched against an identity system to determine if there is any non-trivial manipulation [68]. Popescu and Farid [57] estimated the inter-pixel correlation caused by interpolation for detecting rescaling operations. Farid *et al.* leveraged physics-based properties such as lighting and reflection to identify image forgeries [31, 32]. Ng *et al.* [52] also proposed physics-motivated features to separate realistic photos and computer graphics. Toward a unifying understanding of digital image forensics, a framework of component forensics has been established [70] for the study of more generic scenarios. In parallel to establishing forensic techniques, efforts of *anti forensics* have also been made to examine their vulnerability as well as countermeasures to intentional attacks [64].

In this chapter, we consider the problem of camera model identification that matches digital images against potential models of camera sources. This problem finds its applications in many forensic and homeland security scenarios. For example, a forensic analyst during a crime-scene investigation may find a cell-phone left in the scene. Using existing forensic tools such as the Universal Forensic Extraction Device from CelleBrite Mobile [48], the analyst can extract data from the cell-phone including the user contacts, call history, text messages, and all the images stored on the cell-phone. Among these data, the images taken using the cell-phone's built-in camera potentially sketch what the victim may have seen in his or her last minutes. However, before such images become eligible for forensic evidence, their integrity first has to be established. The analyst can first check whether or not the images are from the exact cell-phone camera that is found, and this can be accomplished using techniques such as the Photo-Response Non-Uniformity (PRNU) [24] that captures the camera-specific characteristics. In the case when the images are *not* from the exact cell-phone camera, it becomes crucial to identify the underlying camera models associated with the images so that the analyst can infer further the images' possible origin.

In the forensic literature, there have been a good number of techniques devoted to camera model identification. One class of techniques approach this goal by identifying the underlying color interpolation algorithm that a digital camera has used to create an image [10, 67]. Color interpolation is a common step in digital photography that has a crucial impact on the quality of resulting images [40]. As different camera manufacturers compete with customized color interpolation algorithms to enhance visual quality, it has been shown that the make and model of the source camera can be inferred from the underlying color interpolation algorithm [10, 17, 67]. While promising results have demonstrated the effectiveness of this approach, we show in this chapter that the achievable identification performance has a substantial dependency on the types of image content. Based on the scheme proposed in [67], we provide a detailed investigation of such content dependency. Both experimental and analytical studies suggest that the image-extracted color interpolation parameters have different statistical distributions with respect to image content. As a result, image content plays a role in the achievable identification performance, and the performance can be penalized if there exists mismatch between the content of images used during training and testing phases. Such an understanding can not only provide a rule of thumb for manually selecting proper training images, but also lead to automatic training image selection schemes proposed in this chapter. By automatically incorporating content awareness into the selection of training images, we can save the workload of tedious training image selection saved, and improve the identification performance for both seen and unseen image content. Finally, as content dependency is an inherent issue that can occur in other identification schemes, we expect that the proposed content-aware methodology will have a broader impact and more upcoming applications.

The rest of the chapter is organized as follows. Section 2.2 reviews the basics of camera model identification based on the traces of color interpolation. Section 2.3 investigates the content dependency of camera model identification. The developed understanding of content dependency is then applied in Section 2.4 to implement content-aware selection of training images. Section 2.5 proposes the profile-based adaptive training to further exploits content awareness. Section 2.6 considers the extension of the proposed selection schemes to other types of image content. Section 4.7 summarizes this chapter.

2.2 Camera Model Identification using Color Interpolation Traces

2.2.1 Color Interpolation in Digital Imaging Pipeline

Most digital cameras in today's consumer market follow a similar imaging pipeline as illustrated in Fig. 2.1. Light reflected from the real-world scene passes through the optical components and is then detected by an array of sensors. As the sensors are only capable of detecting the light intensity, in order to acquire color information, a color filter array (CFA) is employed to filter the lights and selectively allows a certain color component of light, commonly either red, green, or blue, to reach the sensors. A predetermined CFA pattern dictates what color component is allowed to pass at each sensor, and this pattern contains usually a periodic repetition of the 2×2 Bayer pattern or its shifted variants shown in Fig. 2.2. Once the data obtained from the CFA is available, the intermediate pixel values lost in color sensing are interpolated using its neighboring pixel values by an operation commonly known as *color interpolation* or *demosaicing*. Following color interpolation is a post-processing stage, in which various types of in-camera processing operations such as white balancing, gamma correction, and compression may be performed to enhance the overall picture quality and/or to reduce storage demand. The result of the post-processing stage is the final camera output.

Since a substantial amount of color information is lost in terms of spatial resolution during color acquisition, color interpolation has a crucial impact on the quality of final image outputs [40] and has been an active research area in image processing. Detailed surveys and comparisons of color interpolation techniques can

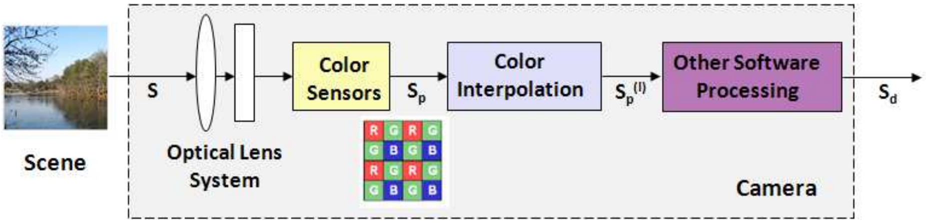


Figure 2.1: The Imaging Model Inside Digital Cameras.

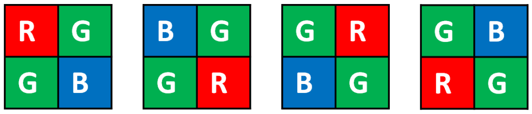


Figure 2.2: Bayer pattern and its shifted variants.

be found in [2, 40]. The algorithms in the literature range from non-adaptive ones with low complexity such as bilinear or bicubic interpolation to highly adaptive and complex ones that can better capture the underlying image structure and recover the lost color information. Different camera manufacturers customize color interpolation algorithms to enhance visual quality, and therefore it has been found that the source camera make and model can be effectively identified by first determining the underlying color interpolation algorithm [10, 67].

2.2.2 Existing Identification Schemes

A few prior works have studied how to identify the underlying color interpolation algorithm of a camera-generated image [5, 10, 57, 67]. In a nutshell, these works consider different parametric models that can characterize a variety of color interpolation algorithms, and the parameters associated with a particular algorithm are estimated using sample images processed by the algorithm. These parametric models differ in their trade-offs between flexibility and complexity, namely, how well they can approximate a given color interpolation algorithm versus how much data is needed for parameter estimation. The works in [5,57] use expectation-maximization (EM) techniques to compute a set of weights for classifying several color interpolation algorithms. Quadratic pixel correlation coefficients are employed in [41] as color interpolation traces for camera model identification. The work in [67] proposes a region-wise linear interpolation framework in which pixels are grouped into different directional regions and pixels belonging to the same region share the same linear interpolation. The CFA pattern and the linear interpolation coefficients associated with each region can be jointly estimated using least-squares methods. In [10], a partial derivative correlation model is introduced to incorporate the higher-order relation among pixels as well as the cross-color channel correlations that are not explicitly addressed in [67]. The parameters of this model can also be estimated by an EM algorithm [10].

Among these existing works, the scheme proposed by Swaminathan *et al.* [67] is one of the earliest that incorporates the concept of direction-adaptive interpolation and has been shown to have a promising identification performance [17]. This scheme has also been used as a building block to regularize the behavior of blind image deconvolution using the color interpolation regularity [68]. Despite the promising identification accuracy reported in previous works, we shall show in this chapter that the achievable accuracy has a substantial dependency on the content of the sample images. We investigate such content dependency through both experimental studies and analytical justifications, and demonstrate how the identification performance can be improved by properly incorporating the content dependency into identifier design.

2.2.3 Refined Color Interpolation Identification Scheme based on Swaminathan *et al.* [67]

To study the content dependency of color interpolation identification, we implement the identification scheme proposed by Swaminathan *et al.* [67], which is one of the earliest works that incorporates the concept of direction-adaptive interpolation and has been shown to have a promising identification performance. To better reflect the state of the art, we improve this scheme by refining its directional classification rules for higher identification accuracy. Specifically, let $I_{x,y}$ represent the sensor value at location (x, y). The local gradient profile along different directions can be found as:

$$\begin{cases} H_{x,y} &= |I_{x,y-2} + I_{x,y+2} - I_{x,y}|, \\ V_{x,y} &= |I_{x-2,y} + I_{x+2,y} - I_{x,y}|, \\ D_{x,y} &= |I_{x-2,y-2} + I_{x+2,y+2} - I_{x,y}|, \\ A_{x,y} &= |I_{x-2,y+2} + I_{x+2,y-2} - I_{x,y}|. \end{cases}$$

Each pixel at location (x, y) is classified into one of five directional regions that are preset using two thresholds T_1 and T_2. As illustrated in Fig. 2.3, Region R_1 contains pixels satisfying $H_{x,y} - V_{x,y} > T_1$, *i.e.*, pixels with a significant horizontal gradient; Region R_2 has pixels satisfying $V_{x,y} - H_{x,y} > T_1$, *i.e.*, pixels with a significant vertical gradient. Similarly, Region R_3 contains pixels with a significant anti-diagonal gradient satisfying $A_{x,y} - D_{x,y} > T_2$, and Region R_4 contains pixels with a significant diagonal gradient satisfying $D_{x,y} - A_{x,y} > T_2$. Pixels not in any of the above are assigned to Region R_5, which mainly come from smooth areas.

With a given CFA pattern, the set of locations in each color channel that are acquired directly from the sensor array can be determined. By approximating the remaining pixels to be interpolated with a set of linear equations in terms of the colors of directly-captured pixels, we can obtain a set of linear equations corresponding to each directional region (R_1, R_2, R_3, R_4, R_5) in each color channel (red, green, and blue). Let each set of equations for a particular directional region and color channel be represented by

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{2.1}$$

This set of equations can be solved for the linear interpolation coefficients and the resulting interpolation error using the least-squares method. Specifically, the least-squares solution to the above equation set is given by

$$\hat{\mathbf{x}} = \mathbf{A}^{+}\mathbf{b} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

The obtained color interpolation coefficients can then be used to reconstruct the image. For each CFA pattern, one can calculate the reconstruction error, and the optimal CFA pattern and color interpolation coefficients are jointly selected as the combination that yields the lowest reconstruction error.

Thus, from each image, we can obtain a vector of estimated color interpo-

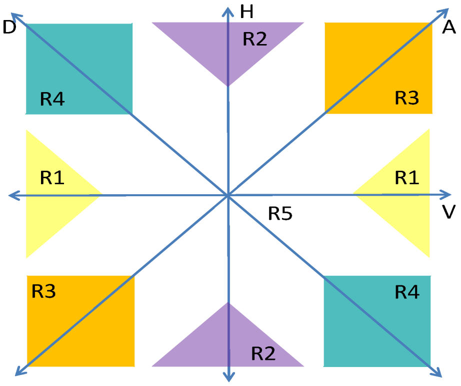


Figure 2.3: Pixel classification based on local gradient.

lation coefficients, which can be subsequently used as features for camera model identification. Although one can apply dimensionality reduction to reduce the feature dimension, we use the estimated color interpolation coefficients as raw features to illustrate some statistical properties that are crucial in this chapter. Finally, machine learning techniques, such as the probabilistic Support Vector Machine [77] adopted in this chapter, can then be employed on the features to construct camera model identifiers.

2.3 Content Dependency of Camera Model Identification

2.3.1 Accuracies of Camera Model Identification under Various Content Conditions

We use 16 different cell-phone camera models listed in Table 5.1 to examine the accuracy of our refined camera model identification scheme. Note that we will use “camera” and “camera model” interchangeably for convenience. It is worth pointing out that a good number of cell-phone cameras are included in this chapter.

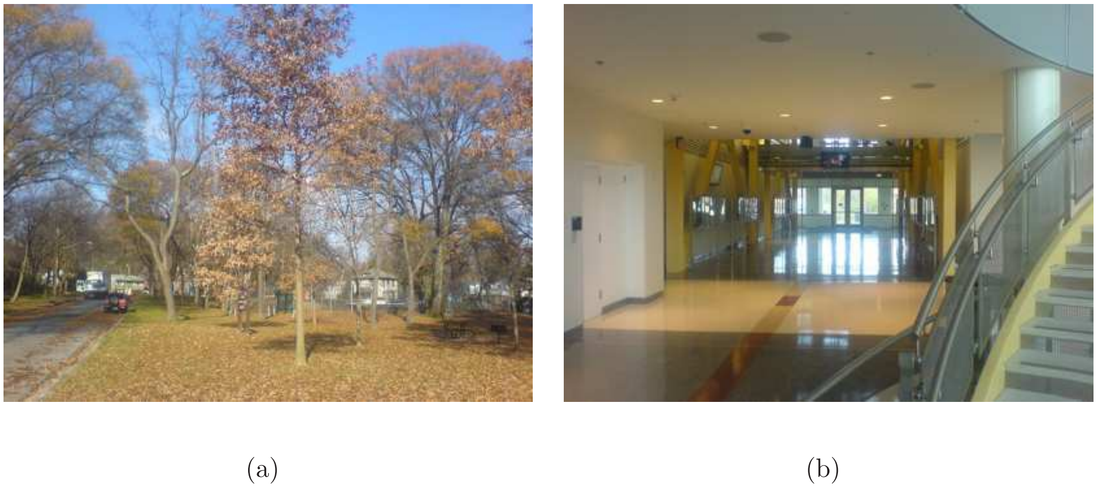
These cameras range from low-end products (for example, Samsung SPH-i700) to more recent releases (for example, Apple iPhone4), and thus our study also sheds light on the camera model identification capability on cell-phone devices in today's consumer market.

With each camera, we have taken 100 images of diverse content as a way to sample the scenes in our environment. These 100 images can be roughly grouped into two types of scene. Fifty images of the first type (called "Type I") are in essence *natural scenes* taken outdoors with substantial texture regions made of natural materials such as trees, leaves, or grass. Fifty images of the other type (called "Type II") are basically *man-made scenes* that contain man-made structures mostly taken indoors. Typical examples of these two types of scenes are shown in Fig. 2.4(a) and 2.4(b), respectively. From each image, a block of 512×512 pixels is extracted from which the color interpolation coefficients are estimated and used as the features for camera model identification. We employ the standard probabilistic SVM with cross validation to train a 16-class camera model identifier [67].

In order to understand the effect of content dependency, we explicitly separate Type I and Type II scenes to form different combinations of training and testing settings and observe the respective identification performances. Fig. 2.5 shows six different training-testing data pairs and the corresponding camera model identification accuracy for different numbers of training image blocks. Note that for the training setting denoted by "Mixture", Type I and Type II scenes are uniformly mixed from which a specified number of training images are selected for training. As we can see, the highest accuracy of around 99.55% is obtained when Type I scenes

Table 2.1: Cell-Phone Camera Models Used for Model Identification Experiment

Index	Camera model	Index	Camera model
1	Sony Ericsson W810i	9	Samsung SCH-i760
2	Sony Ericsson W760a	10	Samsung A707
3	Sony Ericsson W705a	11	Samsung SPH-i700
4	LG VX-9700	12	Nokia E71x
5	LG VX-8550	13	Nokia 6650d
6	LG VX-8100	14	Blackberry Bold 9700
7	Apple iPhone 3G	15	Motorola Cliq
8	Apple iPhone 4	16	HTC Apache



(a) (b)

Figure 2.4: A typical Type I scene (a) and a typical Type II scene (b) in our experiment.

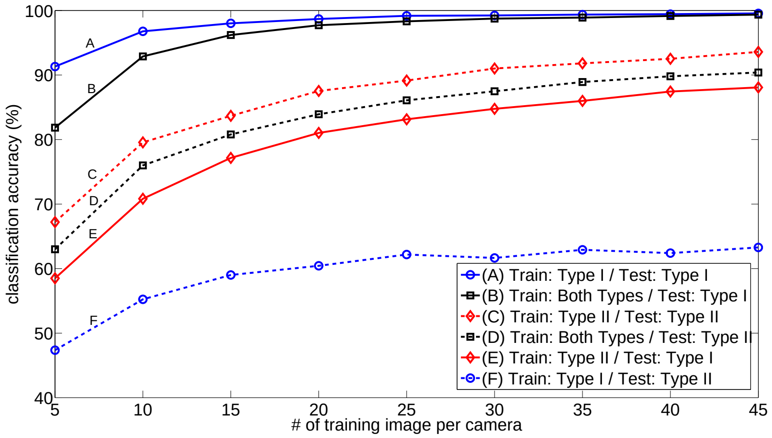


Figure 2.5: Testing accuracies for different combinations of training / testing data.

are used both for training and testing. However, the accuracy drops drastically to only 63.28% as we use a classifier trained with Type I scenes for training and test with Type II scenes. If we instead use Type II scenes for training, we can obtain an accuracy of 88.08% when Type I scenes are tested and 93.58% when Type II scenes are tested, respectively. Such a trend is essentially consistent as the number of training images grows.

The identification results shown in Fig. 2.5 have two implications. First, the identification accuracy is penalized if the training image data and the testing image data do not match in terms of their *content*. Second, camera model identification using Type II scene images appears to be more difficult compared to using Type I scene images. It is of interest to investigate the underlying reasons for these findings, to which we take a statistical approach in the next subsection.

2.3.2 Distributions of Coefficient Estimates Associated with Different Scenes

With each image block, we obtain a vector consisting of the estimates for the color interpolation coefficients, and study the distribution of the coefficient estimates due to content dependency. We calculate the statistics of the coefficients estimates, including the mean and variance, in order to understand if Type I and II scenes lead to distinct distributions.

We first compare the mean coefficient estimates associated with Type I and II scenes using the two-sample t-test [19], which is a statistical tool for determining if two sets have different mean values. For two given sets of one-dimensional samples $\{x_i\}$ and $\{y_i\}$, the hypothesis test can be formulated as

$$\begin{aligned} H_0 : \bar{x}_i &= \bar{y}_i \\ H_1 : \bar{x}_i &\neq \bar{y}_i \end{aligned}$$

with respect to a given *significance level*, *i.e.*, a probability threshold below which H_0 will be rejected. For multi-dimensional samples, we define the *strength* of mean-value difference as the percentage of dimensions on which the hypothesis H_0 is rejected (*i.e.*, the two sets have distinct mean values over the particular dimension). For a significance level of 0.05, we show in Fig. 2.6 the strength of mean-value difference between opposite types of scene (that is, Type I versus II), as well as the strength of mean-value difference between complementary subsets of the same type of scene, repeated for each camera. It can be seen that the strength of mean-value difference is consistently larger when we consider coefficient estimates from opposite types

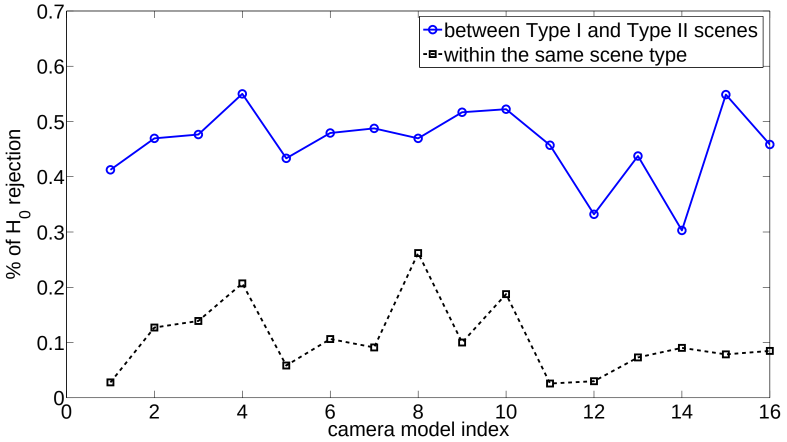


Figure 2.6: Strength of mean-value difference of coefficient estimates per camera. For multi-dimensional features, the strength is defined as the percentage of dimensions on which H_0 is rejected, *i.e.*, the mean values are distinct over the particular dimension.

of scenes, suggesting that the two types of scenes have unequal mean coefficient estimates.

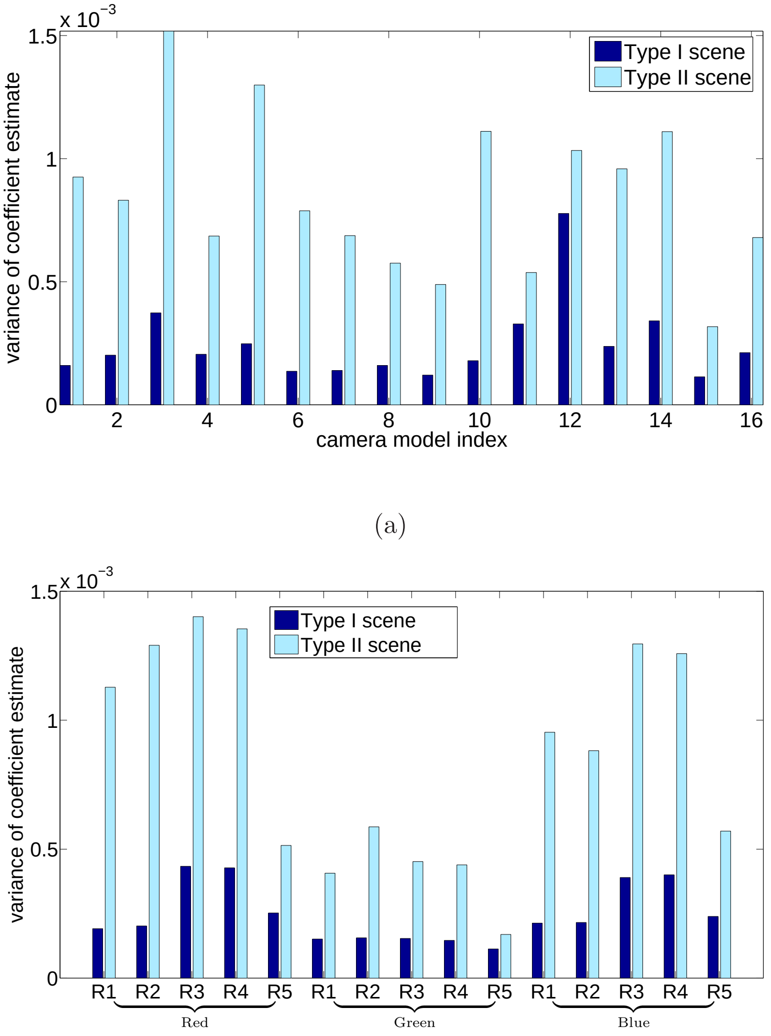
We also examine the variance of coefficient estimates associated with Type I and II scenes. For each camera, we average the variance of coefficient estimates over each dimension, which is shown in Fig. 2.7(a). It can be seen that the variance of coefficient estimates associated with Type I and II scenes are significantly different. In particular, for each camera under consideration, the coefficient estimates from Type I scenes have a smaller variance than those from Type II scenes. Such a difference can also be observed by calculating the variance associated with individual directional regions (R_1 to R_5) in Type I and II scenes, as shown in Fig. 2.7(b). The variance associated with Type I scenes is consistently lower in all directional regions with a clear margin.

2.3.3 Impact of Characteristics of Type I and II Scenes on Coefficient Estimation

Once we see the differences in the mean and variance of coefficient estimates associated with different scenes, it is of interest to obtain a deeper understanding of such differences as well as their impacts. We explore here the fundamental characteristics of Type I and II scenes that lead to such differences, and how the identification of camera model identification is impacted.

The difference in the mean coefficient estimate can be attributed to estimation bias, which arises as an adaptive color interpolation is approximated by the directional linear interpolation model. The ad-hoc partitioning of pixels into a fixed number of directional regions may not perfectly match the underlying interpolation algorithm, and hence each directional region may contain pixels that fit the region to different extents. Such limited fitting of directional regions makes the coefficient estimates biased, and the extent of bias depends on the overall composition of pixels, which is controlled by the scene type of images. This suggests the difference in mean coefficient estimates associated with different scenes, and is consistent with our result of the two-sample t-test.

The difference in the variance of coefficient estimates can be understood as follows. After all pixels are assigned into one of the directional regions, each pixel contributes to its directional region a linear equation that encodes the relation be-



(a)

(b)

Figure 2.7: (a) Variance of coefficient estimates per camera. (b) Variance of coefficient estimates per directional region.

tween the pixel and its neighbors, in terms of the color interpolation coefficients. For each directional region R in each color channel C, the variance of the coefficient estimate is determined by two factors: the variability of the solution when an individual equation is solved, denoted by $\sigma^2(R, C)$, as well as the number of equations available for each directional region, denoted by $N(R, C)$. $N(R, C)$ can be directly calculated, whose average values for different (R, C) are plotted in Fig. 2.8. We can see that in Type I scene images, much more pixels are assigned into R_1 to R_4. To estimate $\sigma^2(R, C)$, we calculate the variance of coefficient estimates when an upper threshold is placed for the number of equations used for coefficient estimation. Note that the exact number of equations used in the estimation can be smaller than the threshold, but is equal or close to the threshold when the threshold is small. For illustration, we plot in Fig. 2.9 this variance with respect to different thresholds for two directional regions, R_1 and R_5 of the red channel. It can be seen that: 1) the variance decreases with respect to the threshold; 2) for the same number of equations, which coincides with smaller thresholds, coefficient estimates from Type I scenes have lower variance. This holds regardless of the directional region, and implies that $\sigma^2(R, C)$ associated with Type I scenes is larger than that associated with Type II scenes. Furthermore, the effect of $\sigma^2(R, C)$ is more dominant than that of $N(R, C)$. In particular, although more equations are available in R_5 of Type II scenes, the variance associated R_5 in Type I scenes is still lower.

The difference in $N(R, C)$ and $\sigma^2(R, C)$ between Type I and II scenes can be attributed to the fundamental difference of these scenes in terms of the gradient distributions shown in Fig. 2.10. Recall that Type I scenes are essentially natural

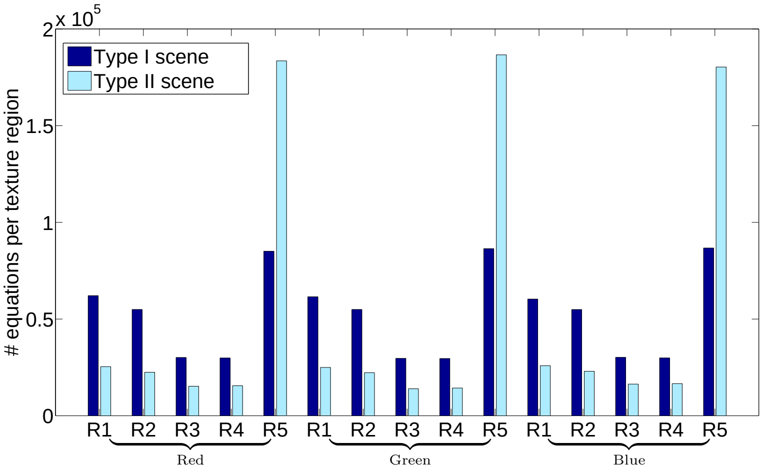
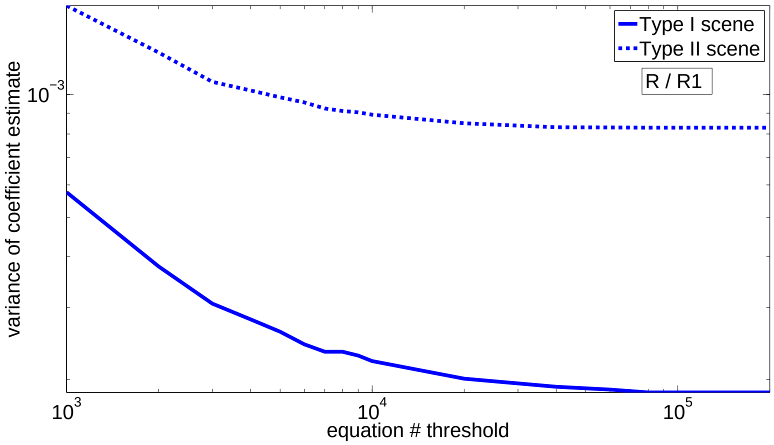


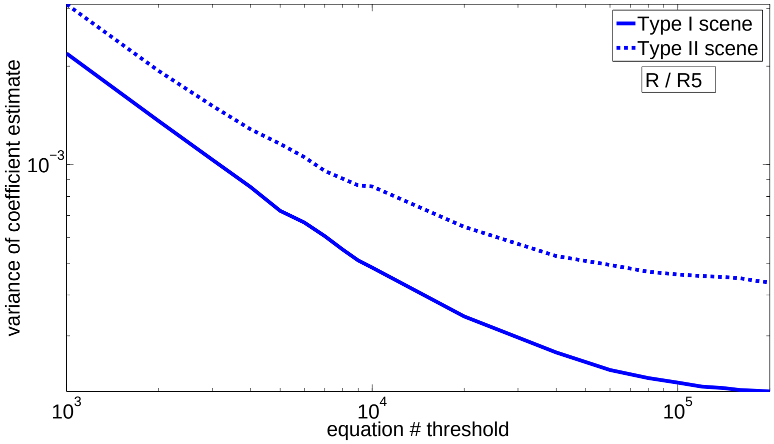
Figure 2.8: Number of equations per directional region.

scenes while Type II scenes are essentially man-made structures (see Fig. 2.4). The gradient of Type I scenes has more large values compared to that of the Type II scenes, which is expected since Type II scenes have significant portions of smooth areas without large variations. Consequently, more pixels in Type I images will be assigned to R_i $(i = 1, 2, 3, 4)$, leading to larger $N(R, C)$. On the other hand, larger pixel-value variations in Type I scenes impose more constraints on the inter-pixel relations; therefore, individual equations have more consistent solutions and thus $\sigma^2(R, C)$ is smaller. One effective measurement of the solution's consistency of the equations is the *condition number* [76]. A widely adopted definition of the condition number is the ratio of the maximal singular value of the matrix $\mathbf{A}$ in (2.1) to the minimal one. The smaller the condition number is, the more consistent a solution is. We plot for illustration the average condition number with respect to different equation number thresholds for R_1 and R_5 in the red channel, in Fig. 2.11, where the substantial margin between Type I and II scenes confirms the aforementioned difference in solution consistency.

Our findings above suggest that coefficient estimates associated with Type I and II scenes have unequal values of mean and variance. In other words, Type I and II scenes lead to different distributions of coefficient estimates. A direct consequence of this difference is that the identification accuracy will be penalized if the content of the training and the testing images do not match. Furthermore, the substantial difference in the variance associated with Type I and Type II scenes explains why the identification accuracy is lower when Type II scenes are used. One way to predict the identification performance is to jointly estimate the *between-camera scatter*, which is defined as the average difference between mean coefficient estimates associated with individual cameras, and the *within-camera scatter*, which is represented by the average variance of the coefficient estimates associated with individual cameras. Whereas the between-camera scatters for Type I and II scenes are close to each other and differ by only 5%, the within-camera scatter for Type I scenes is consistently and substantially lower than that for Type II scenes, as shown in Fig. 2.7. In other words, the coefficient estimates associated with Type I scenes are more consistent and form a denser distribution, making it easier to distinguish different cameras when only Type I scene images are considered. Conversely, estimates associated with Type II scenes are less consistent and spread more widely; they are more likely to overlap with each other and thus the camera distinguishability is lower.



(a)



(b)

Figure 2.9: (a) Variance of coefficient estimates with respect to the equation number threshold (R_1 in red channel). (b) Variance of coefficient estimates with respect to the equation number threshold (R_5 in red channel).

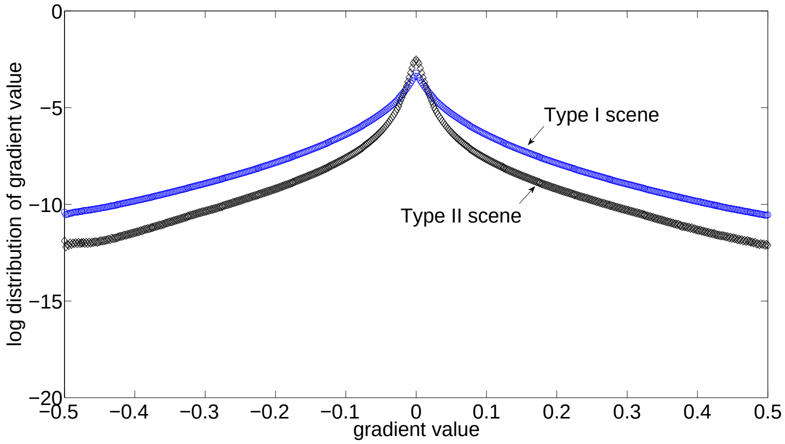
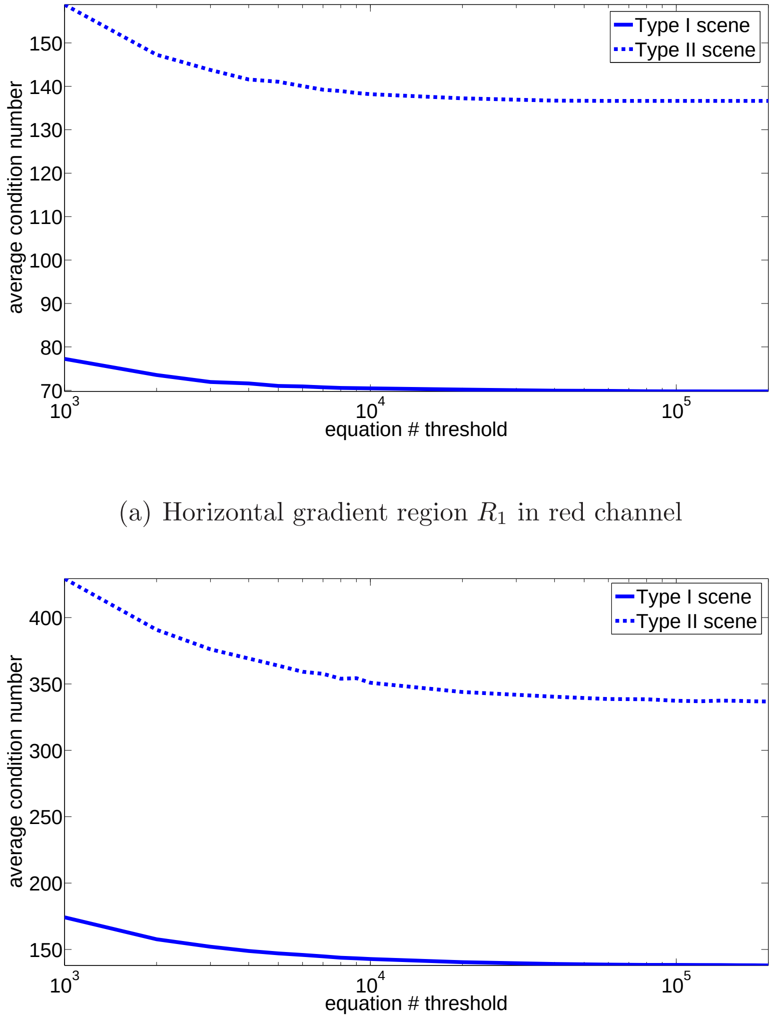


Figure 2.10: Gradient distributions of Type I and Type II scenes.

2.4 Content-Aware Selection of Training Images

2.4.1 Semi Non-Intrusive Training for Completely Non-Intrusive Testing

Our investigation in Section 2.3 suggests the substantial content dependency of camera model identification, and such dependency can degrade the achievable identification performance. In many forensic scenarios involving camera model identification, the analyst has no control over the images to be matched against a target camera, but with the camera at hand, he/she is able to specify the training process. Specifically, the forensic analyst is provided with the extra freedom of generating and selecting training images that match the testing image, so as to mitigate the content mismatch problem. Note that in reality, it may be difficult to evaluate certain quantitative properties of a scene during the collection process of training images. That is, the image collector may be unable to decide if a scene matches the



(a) Horizontal gradient region R_1 in red channel

(b) Smooth region R_5 in red channel

Figure 2.11: (a) Average condition number with respect to the equation number threshold for R_1 in red channel; (b) Average condition number with respect to the equation number threshold for R_5 in red channel.

testing images using a given quantitative measure, unless some built-in functionalities or feedback channels via the network infrastructure are available. Alternatively, we assume in this chapter that a "super set" of training images is first collected without full awareness of the image content (the rule of thumb above is still useful). Proper training images that are tailored to the testing images are then selected offline. In order to capture the variations of coefficient estimates, Section 2.3 shows that training data should include a sufficient number of Type II scenes. In addition, a small number of Type I images should also be included so that the camera model can be accurately identified using images of the Type I scene.

2.4.2 Fitness Evaluation of Training Images

The aforementioned finding can be used as a rule of thumb to guide the collection process of training image data. In reality, however, a hard division of image content into Type I and II scenes is not always straightforward, and ambiguity can easily arise for images with mixed content. [56]. More than that, it is a heavy burden to manually select training images that belong to a particular type of scene. In view of these two reasons, it is desirable to avoid the hard and manual division of training images.

In order to select training images automatically in terms of their content characteristics, we need to define: 1) image representations that stand for the image content and 2) quantitative measures that evaluate the similarity between two content representations. As discussed in Section 2.3.2, training images should match the statistical distribution of the testing images. However, properties such as mean and variance are the ensemble statistics calculated over multiple images, and cannot be directly obtained with individual ones. Nevertheless, our observations made in Section 2.3.2 suggest several possible "profiles" that can be immediately extracted from each image to represent their content. Specifically, we propose and examine the region partitioning profile (RPP) and the condition number profile (CNP). The RPP is defined as the concatenation of the numbers of pixels that are grouped into the 15 directional regions (R_1 to R_5 in red, green, and blue channels). The CNP is defined similarly as the concatenation of the condition numbers associated with the 15 directional regions. Clearly, these two profiles can be evaluated from each individual image. Once these profiles are defined, we adopt in this chapter the Euclidean distance between two profiles as a measure of the content dissimilarity between two images.

2.4.3 Selection Strategies

We examine here if the accuracy of camera model identification can be improved by incorporating the content awareness of training images. Toward this end, we consider several different settings that correspond to different levels of content awareness:

- *Blind Content Selection*: Fifty Type I and fifty Type II scene images from each camera are mixed into the super set of training images. A subset of training images is blindly selected and used to construct a camera model identifier.

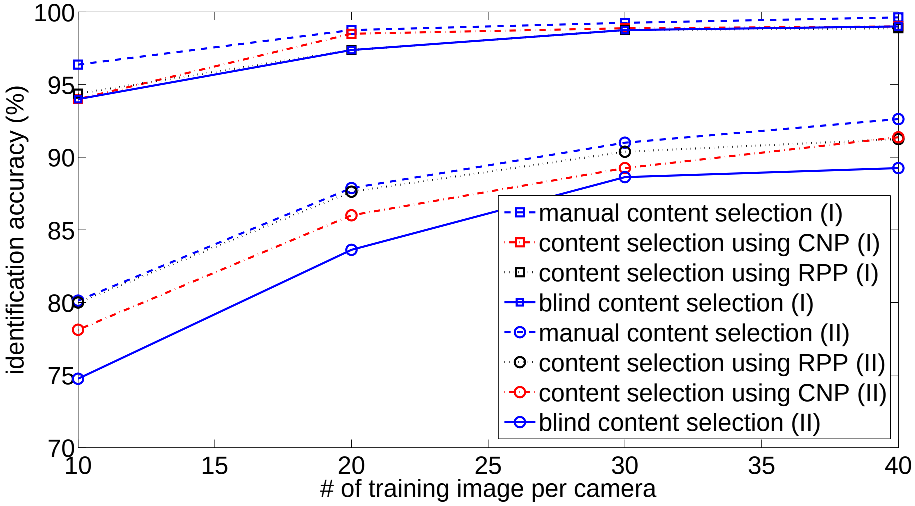
- *Manual Content Selection*: Training images in the super set are classified manually as Type I and Type II ones. Two camera model identifiers are then constructed using the Type I and Type II scene images, respectively. The same manual content classification is also conducted for the testing image and the appropriate camera model identifier is selected accordingly.

- *Automatic Content Clustering Using the Proposed Profiles*: Using either the RPP or CNP to represent a training image, we calculate the Euclidean distance between any two training images as their dissimilarity. A K-means clustering procedure for 2-class is then conducted over the entire super set to automatically partition the training images into two clusters. We expect that the two clusters correspond to Type I and Type II scene images, respectively.

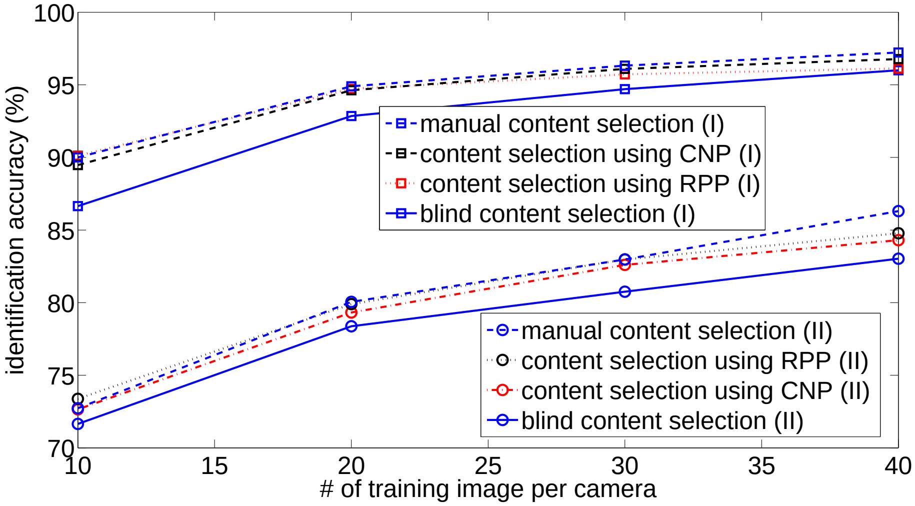
We use two image sets as the super sets to examine various strategies for training image selection. In addition to the set generated by the 16 cell-phone cameras as listed in Section 2.3.1, we create a super set that consists of images that have been explicitly color interpolated using 8 different interpolation algorithms: the first six are well known algorithms, including bilinear, bicubic, smooth hue, median filter based, gradient based, and an adaptive color plane algorithm [2]. In recent years, significant progress has been made to improve the reconstruction quality of color interpolation. To reflect the advancement of the state of the art, we also include a recent algorithm based on local polynomial approximation (LPA) and intersection of confidence intervals (ICI) [55], which performs well in a comparative survey [40], and a latest algorithm that combines local directional interpolation

(LDI) and nonlocal adaptive thresholding (NAT) [80]. The same composition of Type I and Type II scene images, namely 50 Type I scene and 50 Type II scene images, are then included to form the second super set. Incorporating this extra super set is meant to further validate the effectiveness of the proposed concept of content awareness. Using images with synthetic color interpolation also makes it more feasible to expand the scope of content.

Fig. 2.12(a) and 2.12(b) show respectively the identification accuracies for the two super sets. We explicitly separate Type I (I) and Type II (II) scenes to inspect the effectiveness with respect to particular image content. We can see that blind content selection always yields the lower accuracy, which suggests the importance of content awareness. Blind selection can become even less accurate if the training images are mixed in an unfavorable manner. For example, if only 1/5 of the images in the first super set match the testing image, then the identification accuracy for 10 training images per camera drops from 84% to 74%. In the meantime, both manual and automatic content selection using either RPP or CNP outperform blind content selection with similar accuracy improvements. That is, our proposed automatic content selection can effectively replace the tedious manual selection process without sacrificing the identification accuracy. The results also suggest that the two profiles RPP and CNP can both stand for the image content and have comparable performances.



(a)



(b)

Figure 2.12: Comparison of content selection schemes using (a) camera-generated image data; (b) image data with synthetic color interpolation.

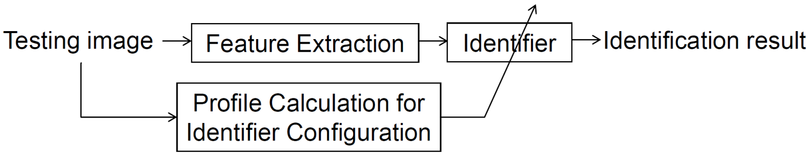


Figure 2.13: Profile-based adaptive training scheme.

2.5 Profile-based Adaptive Training

In this section, we further exploit the notion of content awareness to improve the accuracy of camera model identification. We propose a scheme, referred to as *profile-based adaptive training*, whose schematic is shown in Fig. 2.13. The basic principle of this scheme is to configure the camera model identifier according to the profile of each testing image, so that the resulting identifier better matches the characteristics of the testing image. We consider a special version of profile-based adaptive training, which aims at selecting a given number of training images for each camera model from the super set. The selected training images are then used to train a camera model identifier implemented by a learning algorithm, such as the SVM employed in this chapter. While the manual and automatic content selection schemes discussed in the previous section can be viewed as *non-adaptive* training with a fixed number of configurations, this scheme is *adaptive* to each testing image.

This scheme is considered for two main reasons. On one hand, in the forensic circumstance where only a small number of testing images need to be identified, it is feasible to optimize the camera model identifier for each testing image. Such an optimization, *i.e.*, choosing a given number of training images from the super set, leads to a learning process with lower overhead and a lightweight customized identifier. In comparison, learning an identifier using overly many training images from the super set of possibly heterogeneous content may exceed the capacity of the learning algorithm or cause prohibitive time and memory complexities. On the other hand, we expect that such adaptive training can outperform the non-adaptive training strategies, and thus can serve as a better indicator of the achievable accuracy due to content awareness.

2.5.1 Adaptive Training Image Selection via Profile Matching

Following our discussion in Section 2.3, we assume that the set of color interpolation coefficient estimates is a random vector whose distribution is a function of both the camera model as well as the image content. Denote the camera model by c and let the content be indexed by a profile $\mathbf{p}$ (which can be RPP or CNP proposed in this chapter), then the distribution of the coefficient estimate vector $\mathbf{v}$ can written as $\mathcal{D}(\mathbf{v}|\mathbf{p}, c)$. Under our setting, for each candidate camera model c, we assume that we have a super set of training images $\{\mathbf{I}_{c1}, \mathbf{I}_{c2}, \ldots, \mathbf{I}_{cN}\}$, from which we can calculate the corresponding profiles $\{\mathbf{p}_{c1}, \mathbf{p}_{c2}, \ldots, \mathbf{p}_{cN}\}$ and the coefficient estimate vectors $\{\mathbf{v}_{c1}, \mathbf{v}_{c2}, \ldots, \mathbf{v}_{cN}\}$. When a testing image $\mathbf{I}_t$ with profile $\mathbf{p}_t$ and coefficient estimate vector $\mathbf{v}_t$ is given, profile-based adaptive training aims at selecting n_c training images from each camera model c so that the resulting camera model identifier matches the testing image content indexed by $\mathbf{p}_t$; that is, the camera model identifier learns the distribution $\mathcal{D}(\mathbf{v}|\mathbf{p}_t, c)$. Since the training image selection is carried out independently for each camera model, hereafter we omit the camera model c for sake of notational convenience. We also write $\mathcal{D}(\mathbf{v}|\mathbf{p})$ as $\mathcal{D}(\mathbf{p})$ to highlight the mapping from a profile $\mathbf{p}$ to a distribution $\mathcal{D}(\mathbf{v}|\mathbf{p})$.

First assume that $\mathcal{D}_i \triangleq \mathcal{D}(\mathbf{p}_i)$ is available for each $\mathbf{p}_i$, $1 \leq i \leq N$. For a given $\mathbf{p}_t$, profile-based adaptive training selects a subset of indices $\{s_1, s_2, \ldots, s_n\}$ from 1 to N and uses $\{\mathcal{D}_{s_1}, \ldots, \mathcal{D}_{s_n}\}$ to interpolate $\mathcal{D}(\mathbf{p}_t)$. To perform such interpolation, one needs to assume an underlying structure for the mapping $\mathbf{p} \rightarrow \mathcal{D}(\mathbf{p})$ for all the convex combinations of $\{\mathbf{p}_{s_i}\}_{1 \leq i \leq n}$, *i.e.*, $\{\sum_{i=1}^{n} \theta_i \mathbf{p}_{s_i} | \theta_i \geq 0, \sum_{i=1}^{n} \theta_i = 1\}$. For analytical tractability, we assume that the mapping $\mathbf{p} \rightarrow \mathcal{D}(\mathbf{p})$ satisfies $\mathcal{D}(\sum_{i=1}^{n} \theta_i \mathbf{p}_{s_i}) = \sum_{i=1}^{n} \theta_i \mathcal{D}(\mathbf{p}_{s_i})$ for all $\theta_i \geq 0, \sum_{i=1}^{n} \theta_i = 1$. $\mathcal{D}(\mathbf{p}_t)$ can then be optimally determined by expressing $\mathbf{p}_t$ using $\{\mathbf{p}_{s_1}, \ldots, \mathbf{p}_{s_n}\}$ with minimal representation error. If the profile representation error is measured in squared error sense, the subset selection task can be formulated as the following optimization problem:

$$\begin{aligned} \underset{w_1,\ldots,w_N,b_1,\ldots,b_N}{\text{minimize}} \quad & \left\| \sum_{i=1}^{N} w_i b_i \mathbf{p}_i - \mathbf{p}_t \right\|^2 \\ \text{subject to} \quad & b_i \in \{0,1\}, \ \sum_{i=1}^{N} b_i = n, \\ & w_i \geq 0, \ \sum_{i=1}^{N} w_i = 1, \ w_i \leq b_i. \end{aligned} \tag{2.2}$$

In (2.2), variables $\{b_i\}$ are used to specify indices that are selected, and $\{w_i\}$ are weights assigned to selected indices for representing $\mathbf{p}_t$. The constraint $w_i \leq b_i$ is to ensure that if $b_i = 0$ (*i.e.*, if index i is not selected), then $w_i = 0$ as well. The problem (2.2) is difficult to solve primarily due to the multiplicative form of $w_i b_i$ in the objective function and the integer constraints on $\{b_i\}$. To approach this problem, we adopt a two-step relaxation strategy. In the first step, we let the weights $\{w_i\}$ be equally distributed among n selected indices, namely $w_i = 1/n$ if and only if $b_n = 1$. This makes w_i a function of b_i and can be removed from (2.2). In the second step, we relax the constraint $b_i \in \{0, 1\}$ as $0 \leq b_i \leq 1$. After the relaxation, the optimization problem now becomes

$$\begin{aligned} \underset{b_1,\ldots,b_N}{\text{minimize}} \quad & \left\| \frac{1}{n} \sum_{i=1}^{N} b_i \mathbf{p}_i - \mathbf{p}_t \right\|^2 \\ \text{subject to} \quad & 0 \leq b_i \leq 1, \ \sum_{i=1}^{N} b_i = n. \end{aligned} \tag{2.3}$$

which is a quadratic programming (QP) problem and can be solved in polynomial time. Due to the relaxation, the obtained $\{b_i\}$ are not always 0 or 1, although they are usually quite close to 0 or 1 as illustrated in Fig. 2.14. The indices with largest b_i are selected as $\{s_1, \ldots, s_n\}$. Recall that we have assumed $\{\mathcal{D}_{s_1}, \ldots, \mathcal{D}_{s_n}\}$ are available. In reality, we do not have these distributions, but only their realizations $\{\mathbf{v}_{s_1}, \ldots, \mathbf{v}_{s_n}\}$. Nevertheless, we can treat these realizations as approximations of the distributions, and feed them into the subsequent learning algorithm to learn the desired distribution $\mathcal{D}(\mathbf{v}|\mathbf{p}_t)$.

As the QP problem demands non-trivial complexity, an alternative is to simply select $\{s_1, \ldots, s_n\}$ as those that correspond to the n profiles closest to $\mathbf{p}_t$, *i.e.*, profiles with minimum Euclidean distance to $\mathbf{p}_t$. The rationale for this alternative can be understood as follows. When $n = N$, namely when all training images are selected, it can be shown that the solution to (2.2) is $b_i = 1$ and $w_i \propto 1/\|\mathbf{p}_i - \mathbf{p}_t\|$ for all $1 \leq i \leq N$. The quantity $1/\|\mathbf{p}_i - \mathbf{p}_t\|$ stands for the similarity between $\mathbf{p}_i$ and

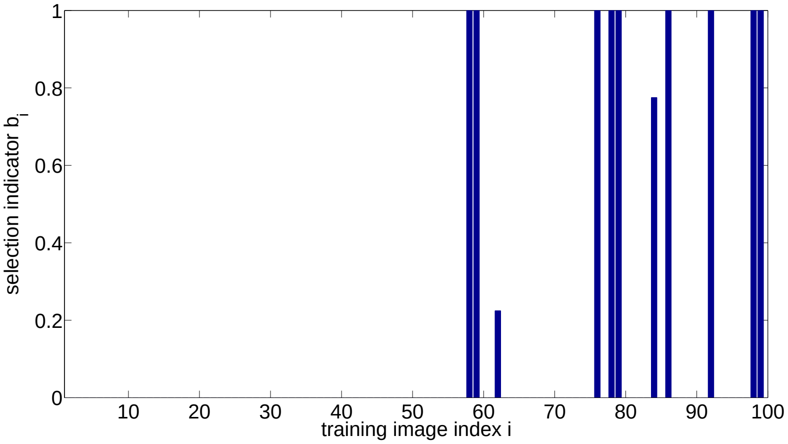


Figure 2.14: A typical solution to (2.3) where $N = 100$ and $n = 10$. Note that most b_is are either 0 or 1.

$\mathbf{p}_t$ and indicates how important each training image is for representing the testing image. Notice that we have implicitly adopted the similarity-based selection in the non-adaptive schemes.

2.5.2 Comparisons and Discussions

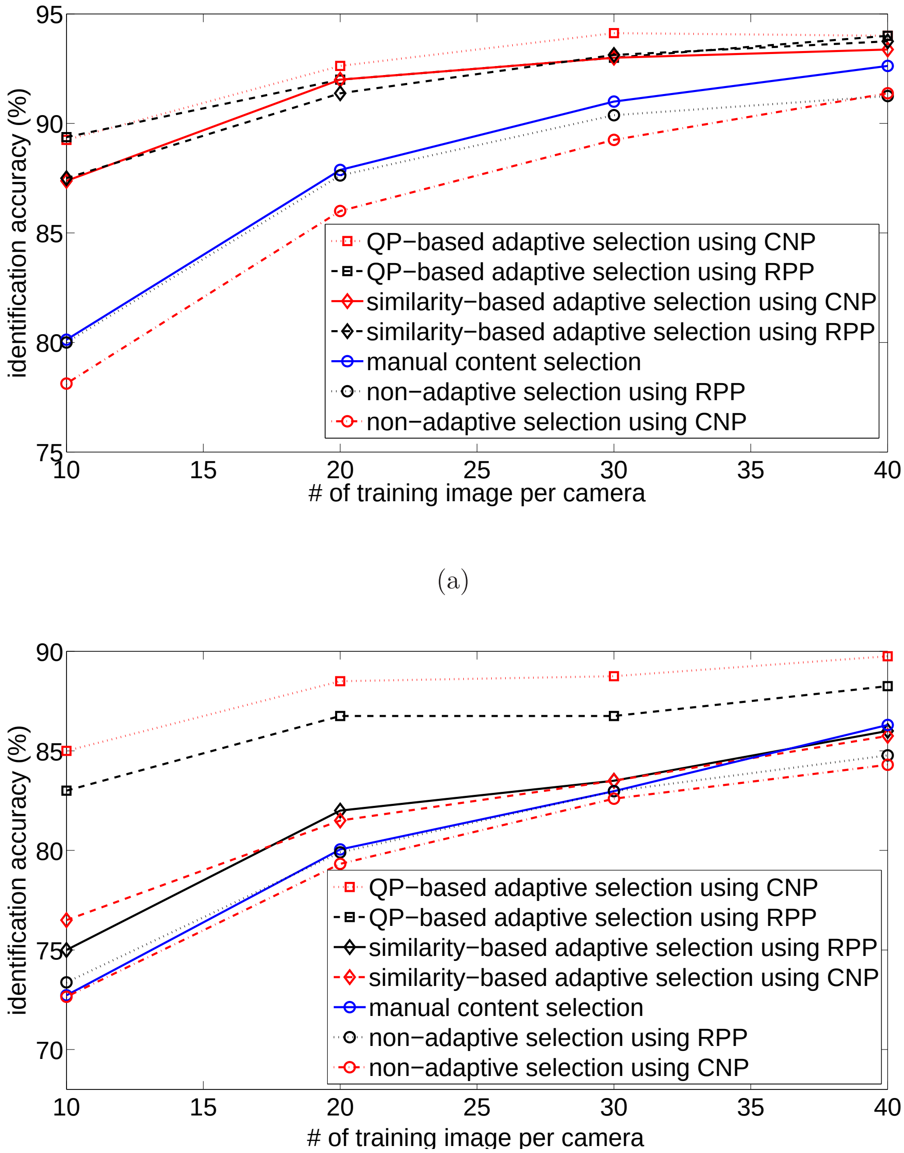
We compare the identification accuracy of the proposed profile-based adaptive training, including both the QP-based scheme and the similarity-based scheme, to those of the content-aware content selection schemes for the two types of image content. Consistent observations are obtained for both types, and we show in Fig. 2.15 the accuracy results over testing images of the Type II scene. We can see that the adaptive selection schemes outperform non-adaptive ones and manual selection, suggesting that optimizing the camera model identifier by adapting to the content of each testing image benefits the identification. Also, the results confirm again the efficacy of the proposed profiles for characterizing the image content. Between the two adaptive schemes, the QP-based one exhibits a substantially higher accuracy, suggesting that the QP solution more accurately approximates the image content using the accessible training images.

Complexity Recall that an important reason for selecting a fixed number of training images via the profile-based adaptive training is to avoid the possible high time and memory overhead when training using the entire super set. When the super set size of each camera model is N, the QP-based scheme that solves (2.3) has time complexity $O(N^3)$ and memory complexity $O(N^2)$, and the similarity-based scheme has time complexity $O(N)$ and memory complexity $O(1)$. Repeating the selection for a total of C camera models requires time complexity of $O(CN^3)$ and $O(CN)$, respectively. While learning a SVM using the entire super set also solves a QP problem [6] and typically requires similar time and memory complexities as solving (2.3), the number of variables in the SVM grows with the number of camera models and thus can incur a substantially higher overhead. For example, if multi-class SVM is constructed in a pairwise fashion [77], then the required time complexity is $\binom{C}{2} O\left((2N)^3\right) = O(C^2 N^3)$, which is higher than both the QP-based and the similarity-based adaptive selection schemes.

Analogies to Other Classifier Adaptation Approaches The proposed profile-based adaptive training builds a camera model identifier that adapts to the content characteristics of each testing image to mitigate the mismatch between the training and testing data distribution. In the literature, such mismatch has also been dealt under the general notion of classifier adaptation within different contexts, such as *domain adaptation* [28] in machine learning and *concept drift* [72] in data mining. Domain adaptation addresses issues such as covariate shift in shared distribution support by reweighting training data samples where the weights are estimated from a bunch of testing data to be classified, and concept drift is handled by maintaining a proper time window that moves over the training data stream for learning the concept and weighting the training data according to age or utility to the targeted concept [36], or by adapting a learnt concept to new training data in an incremental manner without repeated training over used data [78]. We plan to investigate in the future if similar ideas can be incorporated into our profile-based adaptive training. One possible route is to see if we can integrate our training data selection with incremental learning so that a customized identifier can be built by directly adapting an existing identifier to a small amount of training data selected from the super set.

2.6 Extension to Other Image Contents

In previous sections, we have assumed that images can be classified into Type I and Type II scenes. In reality, however, such classification cannot be perfectly definite and a certain ambiguity always exists. In such cases, manual selection of training images may become infeasible, and we resort to our automatic content selection schemes as a possible remedy. In this section, we consider the setting where the same super set



(a)

(b)

Figure 2.15: Comparison of adaptive and non-adaptive content selection schemes (a) camera-generated image data; (b) image data with synthetic color interpolation.

consisting of Type I and Type II training images is collected beforehand, and a separate image set of possibly ambiguous content is used for testing. We conduct two experiments. The first experiment uses composite content with Type I and Type II equally mixed. The second experiment uses another three extra image sets of specialized content.

2.6.1 Composite Content

First, we create a synthetic image set by equally mixing Type I and Type II. More specifically, the left half of each synthetic image is copied from a Type I image, and the right half is from a Type II image. The color interpolation procedure in Section 2.4.2 is carried out to generate eight color interpolated versions of the synthetic image. Such setting mimics the case when a testing image is a composition of Type I and Type II scenes, which can be observed in reality. Under this setting, all the testing images cannot be easily categorized, and therefore it becomes infeasible to manually select the training images. As such, we can only compare the proposed adaptive scheme with blind selection.

As shown in Fig. 2.16, both the similarity-based and the QP-based selection schemes outperform blind selection for images with composite content, and the QP-based scheme is superior to the similarity-based one except when a larger number (> 20) of training images from each camera are used where the two schemes both lead to high ($> 97\%$) identification accuracies. We can see that the identification improvement due to adaptive training and more accurate approximation of testing

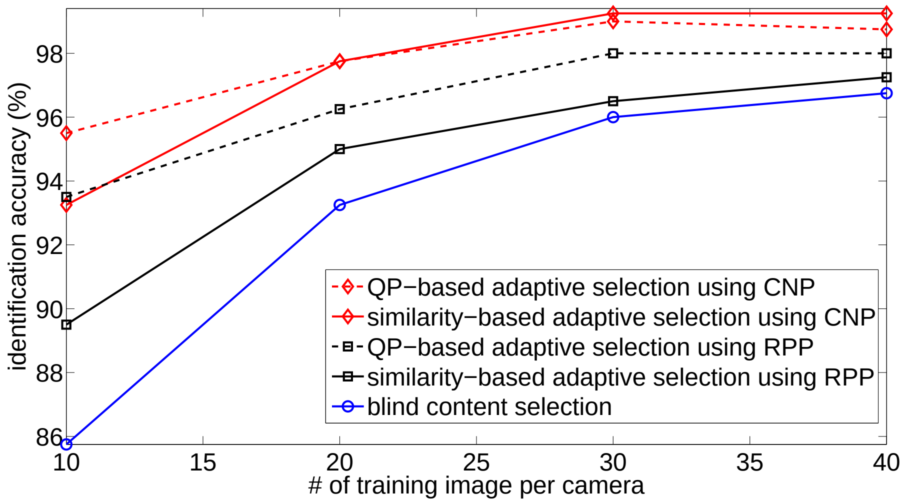


Figure 2.16: Comparison of blind and adaptive content selection schemes for images with composite content.

image content is particularly prominent when the number of training images is smaller. The achievable accuracy for the composite image content is higher than the case of Type II images and slightly lower than the case of Type I images. This is expected since half of each testing image is from Type I. As the image block size is reasonably large (512 × 512 pixels here), there are enough linear equations available for coefficient estimation, and thus the within-camera scatter is small and the identification accuracy is high.

2.6.2 Other Image Contents

We also use three extra sets of synthetic images retrieved on Google Images using keywords “lion”, “sea”, and “texture”, respectively. Examples of these three sets are shown in Fig. 2.17. A closer inspection suggests that the collected “lion”

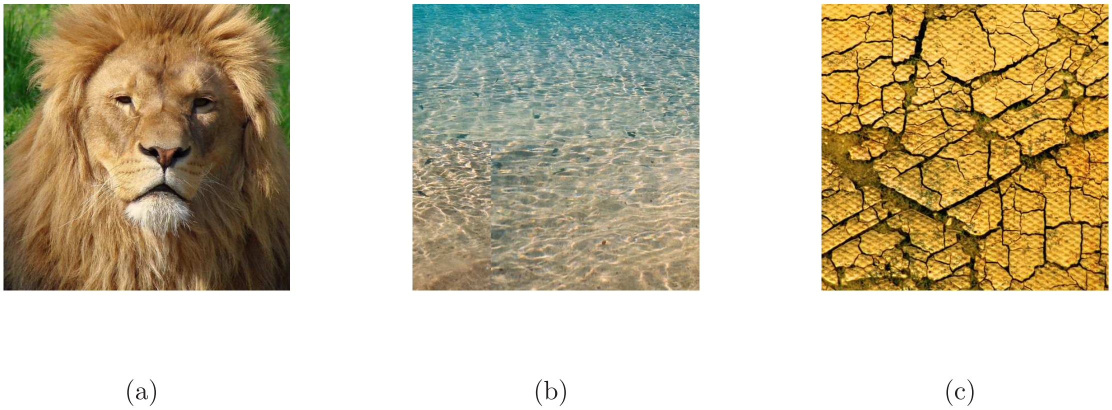
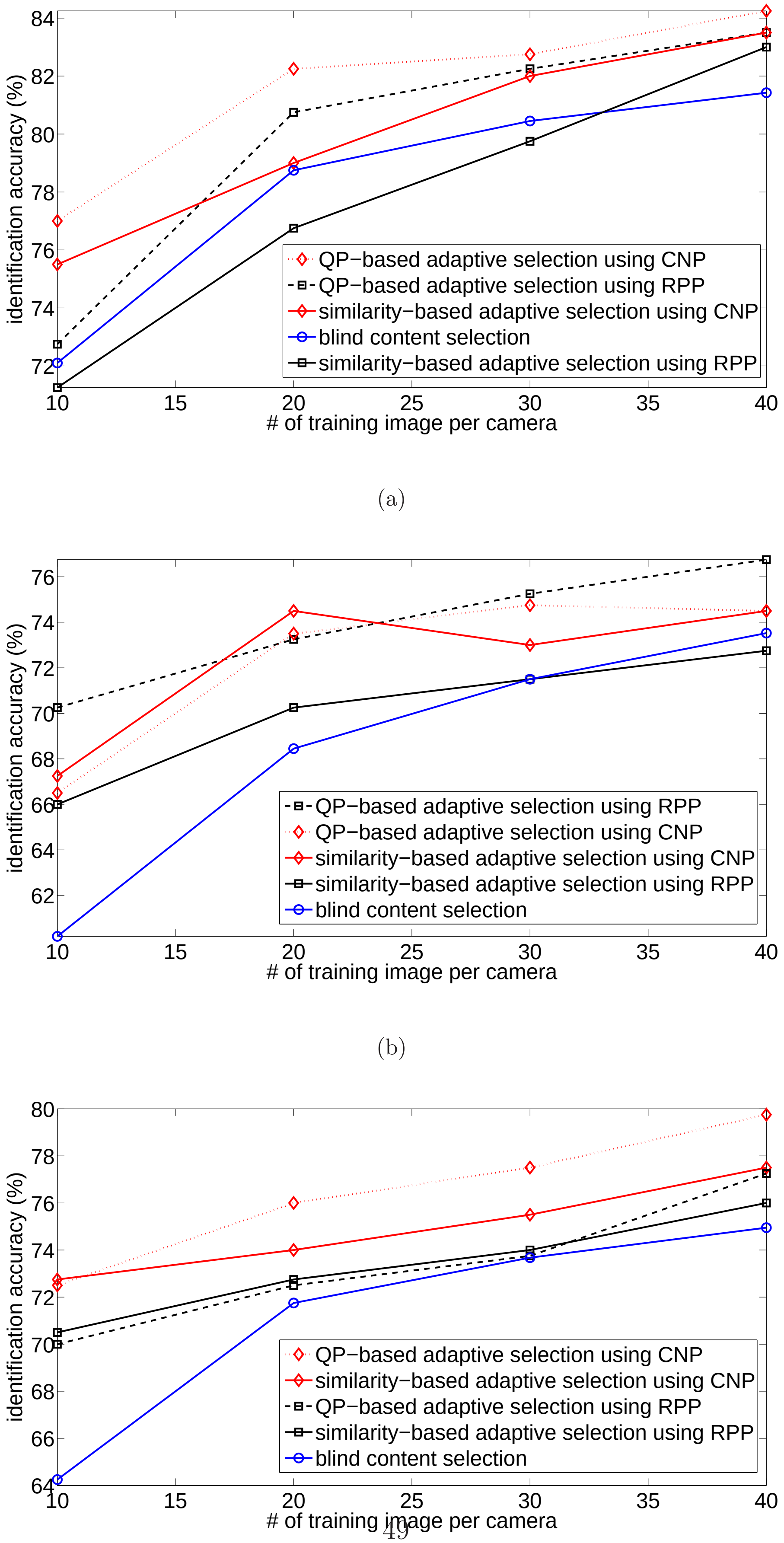


Figure 2.17: Examples of three image categories retrieved from Google Images: (a) lion; (b) sea; (c) texture.

images tend to have textures such as dense hair that shares certain similarity with our Type I images. In comparison, “sea” images are usually smoother and lack rich variations, and “texture” images tend to have more regular variations.

Fig. 2.18 compares blind and adaptive content selection schemes upon the three types of testing images. Except for one case (lion images, blind content selection versus similarity-based adaptive selection using RPP), both adaptive selection schemes outperform blind selection. Also, the QP-based selection scheme leads to more accurate identification than the similarity-based selection scheme, except for the case upon sea images using the CNP profile where the two schemes yield comparable accuracies. We can also notice that CNP seems to be a better representation for these image content categories. Our results here confirm again that the proposed adaptive schemes along with the two profiles can substantially improve the accuracy of camera model identification even for unseen image categories.



(a)

(b)

(c)

Figure 2.18: Comparison of blind and adaptive content selection schemes for (a)

2.7 Chapter Summary

In this chapter, we first present a study of camera model identification using the refined color interpolation coefficient features. Sixteen cell-phone cameras that cover today's consumer market are used for performance assessment. A detailed statistical analysis of the estimated coefficients with respect to different image content shows a substantial content dependency and its impacts on the identification performance. As our study suggests, the image content determines the achievable identification performance, and the identification performance can be penalized due to mismatch between the content of training and testing images. Such an understanding not only serves as a rule of thumb for manually selecting training images that provide sufficient coefficient variations as well as match the testing images, but also leads to automatic training image selection schemes based on our proposed region partitioning profile (RPP) and condition number profile (CNP) that can be easily calculated upon each individual image.

We further propose profile-based adaptive training that can select the optimal training images tailored to the content characteristics of each given testing image. This ensures a lightweight construction of accurate identifiers without incorporating unnecessarily many training images. The selection can be formulated as an profile matching optimization problem that can be relaxed to a quadratic programming (QP) problem and can be solved in polynomial time. Further simplification leads to the selection scheme using the inverse Euclidean distance between two profiles as an indicator for each training image's representation power. As shown in our extensive experiments using both camera-generated and synthetic images, our proposed schemes avoids the tedious manual selection process and significantly improves the identification performance. In particular, when images with content that cannot be easily categorized are tested, our automatic schemes can effectively select the training images systematically and quantitatively.

CHAPTER 3

Camera Model Identification against Anti-Forensics

3.1 Chapter Introduction

Recent years have witnessed a rapid growth of digital imaging technology. The number of pixels on a camera has increased by an order of magnitude in the past decade, and the optical components as well as the signal processing algorithms have also been advanced significantly. Many compact cameras are now equipped with lenses that used to be exclusive for high-end single lens reflex (SLR) cameras, and intelligent in-camera processing modules such as auto focus and color temperature adjustment have become sufficiently reliable to replace manual operations. Most recently, computational photography has begun to impact on how digital image are formed, and new imaging devices such as the light-field camera [51] have emerged in the consumer market as viable options.

As various imaging technologies across different generations are available, new forensic questions about digital images have also been raised and are receiving growing attention. This includes but not limited to: 1) What kind of imaging devices, such as digital cameras, scanners, computer graphics, among others, have been used to create a digital image? 2) If the image is created by, for example, a digital camera, then is it taken by a point and shoot camera, a SLR camera, or a cellphone camera? Further, what is the mostly likely make and model of the source camera? 3) As increasingly more digital cameras now can be equipped with interchangeable lenses, what lens has been used as an image is taken?

To answer these questions, a primary research direction in the literature of *digital image forensics* has focused on the identification of imaging technologies of digital images. One class of techniques addresses the identification of the color interpolation algorithm that a digital camera has used to create an image [5, 10, 57, 67]. Another class studies the classification of source scanners based on noise features [26, 33]. It was investigated in [53] how to differentiate photographic images and computer graphics using physics-motivated properties, and further in [47] how to separate images produced by cameras, scanners, and graphics based on color interpolation and noise statistics features. Recently, more research has been devoted to identifying particular imaging components or imaging characteristics. For example, the identification of SLR lenses was considered in [79], the classification of cellphone cameras was investigated in [11, 17], and the recognition of digital images formed by compressive sensing was discussed in [13].

However, similar to many other tasks regarding data trustworthiness, adversaries who have incentives to perform anti-forensic operations to counter forensic analysis always exist [35, 65]. For example, consider the scenario of technology infringement where a company infringes another company's imaging technology via reverse engineering or industrial espionages. The pirating company has incentives to counteract the identification of color interpolation so that it can use the technology without being caught. It may be of further interest to the pirating company if it can mislead the identification toward a wrong direction that suggests a distinctly different technology. In the scenario of crime scene investigation [48], being aware that information about the source device and the potential owner can be inferred from the imaging technology employed [10, 17, 67], a technology-savvy criminal can conceal the origin of a digital image by circumventing the identification.

These scenarios prompt a strong need for understanding the resilience of today's techniques for identifying digital imaging technologies against anti-forensics. Toward this goal, we have to first explore applicable anti-forensic techniques and evaluate the identification performance against these anti-forensic operations. In principle, one can alter the image to weaken the evidence that may reveal the underlying imaging technology. There exists an inherent trade-off between the strength of the trace concealment and the quality of the resulting image: if the strength is too weak, the identification is likely to remain effective, but if the strength is too strong, the image may suffer from serious distortions. Both situations are unfavorable to the adversary. Different anti-forensic operations may exhibit unequal trade-offs between image quality and identification manipulations; therefore, in order to understand the comprehensive impacts of anti-forensics, it is necessary to examine different options for anti-forensics and compare their trade-offs.

Color interpolation is a commonly used step among various imaging processes involved in today's digital cameras and has a crucial impact on the quality of output images [40]. Different camera manufacturers compete with customized color interpolation modules to enhance the image quality, and it has been shown that the underlying color interpolation method leaves detectable traces in output images that can be leveraged to infer source information such as the camera make and model [10, 17, 67]. In view of the importance of color interpolation identification, we will study in this chapter its resilience against anti-forensic operations, although our methodology is generic in nature and can be easily extended to examine other imaging technologies. To the best of our knowledge, the most relevant work to this chapter is by Kirchner and Böhme in [34], whereby a method was presented to resynthesize a linear color interpolation relation in digital images and minimizes the image quality distortion. Compared to the work in [34], we study counter-identification techniques of lower complexities that can readily applied to a large class of interpolation algorithms that cannot be simply modeled as linear. Our results provide new insights into the resilience of color interpolation identification and reveal inherent vulnerabilities of today's technique. The forensic analyst, once aware of such vulnerabilities, can update the identification technique, which calls for an update on the adversary's side as well. We also formulate such an interplay using a game-theoretic approach and discuss the optimal strategies accessible to a forensic analyst and an adversary.

The rest of the chapter is organized as follows. Section 3.2 reviews color interpolation and its identification based on [67]. Section 3.3 proposes a generic methodology of parameter perturbation for circumventing the identification of a given color interpolation algorithm. Section 3.4 investigates how to mislead the identification toward an incorrect decision. Section 3.5 discusses extensions of the anti-forensic techniques and insights into our study. Section 3.6 summarizes this chapter.

3.2 Design and Evaluation of a Color Interpolation Identification System

In this section, we describe in detail our design and evaluation of a color interpolation identification system, which will be used in subsequent sections for our anti-forensic study.

3.2.1 Mechanism Formulation of Color Interpolation Identification

The fundamental principles and techniques of color interpolation identification as a core element in camera model identification has been explained in details in Chapter 2. We review here some key setups for the sake of self-consistency. In this chapter, we perform the identification of color interpolation based on the scheme proposed in [67]. This scheme is one of the earliest works that incorporates the concept of direction-adaptive interpolation and has been shown to have a promising identification performance. We improve upon the scheme with refined directional classification for higher identification accuracy. Specifically, define $I_{x,y}$ as the sensor value at location (x, y). The local gradient profile along different directions can be found as:

$$\begin{cases} H_{x,y} &= |I_{x,y-2} + I_{x,y+2} - I_{x,y}|, \\ V_{x,y} &= |I_{x-2,y} + I_{x+2,y} - I_{x,y}|, \\ D_{x,y} &= |I_{x-2,y-2} + I_{x+2,y+2} - I_{x,y}|, \\ A_{x,y} &= |I_{x-2,y+2} + I_{x+2,y-2} - I_{x,y}|. \end{cases}$$

Each pixel at location (x, y) is classified into one of five directional regions according to its gradient profile using two preset thresholds T_1 and T_2. The adopted color interpolation identification scheme assumes that pixels belonging to the same directional region are interpolated by a fixed linear interpolation kernel, whose coefficients can be estimated using the least-squares method. The overall color interpolation algorithm can then be represented by a coefficient vector $\boldsymbol{\theta}$ that concatenates all the coefficients associated with each directional region in each color channel.

A general system of identification in our framework learns and matches $\boldsymbol{\theta}$ respectively in a training phase and a testing phase. In the training phase, the forensic analyst learns from some training data the coefficient vector $\boldsymbol{\theta}$ and its possible variations due to the pre-processing and post-process modules. In the testing phase, the forensic analyst matches given testing data against the learnt $\boldsymbol{\theta}$ to determine if they are consistent. Recently, identification of digital devices has been studied more systematically in the context of *component forensics* [70], where different scenarios can be considered depending on the accessibility to the device under question.

Specifically, in the scenario of *intrusive forensics*, the analyst has full access to the device, and can arbitrarily break the device apart to inspect each component inside the device. In the scenario of *semi non-intrusive forensics*, the analyst still has access to the device but cannot break it apart. To build forensic evidence about the components algorithms and parameters, the analyst can only design appropriate inputs to the device and examine the relation between the designed inputs and the corresponding outputs. In the scenario of *completely non-intrusive forensics*, the analyst has no access to the device, and can only use some provided sample device outputs to estimate the component properties. It is clear that these three different scenarios correspond to different levels of forensic capabilities. While the intrusive forensics appears to be very powerful, it may not be always available in reality. Techniques for semi and completely non-intrusive forensics thus may have higher practical values and are the two scenarios of interest in this chapter.

Considering the problem of counter identification based on component forensics, recall for example the infringement detection task described in Section 6.1. Since the owner of the color interpolation technology can select training data to learn the coefficient vector $\boldsymbol{\theta}$, one can assumes that the training phase is (at least) semi non-intrusive. The testing phase is semi non-intrusive if the device made by the pirate company is also accessible to the actual technology owner, and completely non-intrusive if only sample images from the device are available. Without loss of generalizability, we focus in this chapter the combination of a semi non-intrusive training phase and a complete non-intrusive testing phase, and our methodology can be extended to other combinations in a similar fashion. In both phases, an estimate for $\boldsymbol{\theta}$ is obtained from given images. A good number of images are used in the training phase to ensure that the variability due to S' is fully captured, and only a limited number of sample images are available in the testing phase. For the sake of simplicity, we also assume that the processing modules posterior to the color interpolation module is either pre-compensated (for example, if it is known and reversible) or ignorable (if it only introduces minor effects or its effects can be absorbed into color interpolation). We can then formulate the relation between the input scene and output image in terms of the coefficient vector $\boldsymbol{\theta}$ and estimate the distribution of $\boldsymbol{\theta}$ using the training data. Finally, the identification system examines the consistency between $\boldsymbol{\theta}$ estimated during the training and testing phases, and reports an identification confidence $C(I_t)$ of each testing image I_t. More details about these individual steps will be discussed in the following section.

3.2.2 Experiment Setup and Performance Metrics

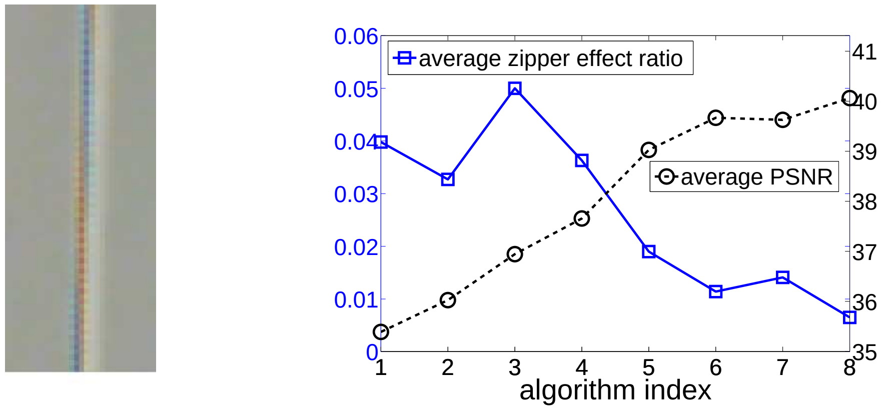
We describe here our experiment setup and performance metrics for carrying out and evaluating anti-forensic schemes. Our goals here are to sample representative color interpolation algorithms used in our study, and to establish a testbed on which we can evaluate forensic and anti-forensic capabilities in terms of identification accuracy and the resulting image quality.

Color Interpolation Algorithms: Color interpolation has been an active research area in image processing. Detailed surveys and comparisons of color interpolation techniques can be found in [2, 40]. The algorithms in the literature range from non-adaptive ones with low complexity such as bilinear or bicubic interpolation to highly adaptive and complex ones that can better capture the underlying image structure and recover the lost color information. We include eight color interpolation algorithms in this chapter. The first six have been well known in the literature for more than one decade, including bilinear, bicubic, smooth hue, median filter based, gradient based, and an adaptive color plane algorithm [2]. In recent years, significant progress has been made to improve the reconstruction quality of color interpolation. To reflect the state of the art, we also include a recent algorithm based on local polynomial approximation (LPA) and intersection of confidence intervals (ICI) [55], which performs well in a comparative survey [40], and a latest algorithm that combines local directional interpolation (LDI) and nonlocal adaptive thresholding (NAT) [80].

We construct a dataset composed of images interpolated by the above eight algorithms. Specifically, we first take 75 high-resolution images with a variety of content by a high-end standalone camera. From each image, we extract the central portion of 1024×1024 pixels, which is prefiltered and down-sampled to 512×512 pixels in order to attenuate the traces of color interpolation and post-processing left by the camera. The resulting 512×512 "full-color" image is then sampled according to a given CFA pattern, and interpolated using each of the eight different interpolation algorithms to simulate in-camera processing.

Performance Metrics: As discussed in Section 6.1, image quality plays an important role in evaluating the performance of anti-forensic operations. We adopt in this chapter the full-reference methodology [75] for image quality assessment whereby the quality of a color interpolated image is assessed with respect to a reference image. The 512×512 full-color image discussed above is used for reference, which is justified in the same way as in [40] and we find that such reference images are visually pleasant. There are a handful of full-reference image quality metrics in the literature. The Peak Signal-to-Noise Ratio (PSNR) is probably the most well-known one. While it is still widely used, previous research has shown that PSNR may not always reflect the true signal fidelity [75]. The quality metric called Structural Similarity (SSIM) index [75] incorporates the similarity in image structure to capture the subjective quality perceived by human beings. One notable artifact in color interpolation is called *zipper effect*, which occurs if an interpolation algorithm fails to interpolate pixels along directional edges, as illustrated in Fig. 3.1(a). The extent of zipper effect can be quantified by the quality metric called *zipper effect ratio* [8, 80] , which measures the increase in spatial color discontinuity due to color interpolation. In order to provide a comprehensive assessment of image quality, it is beneficial to examine more than one quality metric. Fig. 3.1(b) compares the PSNR and the zipper effect ratio of each algorithm, averaging over all testing images. In terms of both metrics, algorithms with higher indices perform better. These algorithms are more sophisticated and represent the advancement of color interpolation technology.

Identification System: We construct a color interpolation identification system that uses the color interpolation coefficients as features. We use the 75 images



(a) (b)

Figure 3.1: (a) An example of zipper effect (best viewed on a screen); (b) PSNR and zipper effect ratio averaged over 50 images associated with different interpolation algorithms: (1) bilinear, (2) bicubic, (3) smooth hue, (4) median filter based, (5) gradient based, (6) adaptive color plane, (7) LPA-ICI, and (8) LDI-NAT.

described above and their interpolated versions created by each of the eight interpolation algorithms. The total number of interpolated images is therefore $75 \times 8 = 600$. Totally 400 of these images are used for training an 8-class probabilistic Support Vector Machine (pSVM) classifier [67] with parameters selected by cross validation, and the remaining 200 images are used for testing. The identification system takes an image as input, and outputs the likelihood of each of the eight algorithms. Maximum-likelihood classification yields an overall accuracy of 96.3%, suggesting the accuracy of color interpolation identification. The maximum likelihood is then adopted as the identification confidence of the classification result.

3.3 Circumventing Color Interpolation Identification via Parameter Perturbation

Our first anti-forensic goal is to circumvent the identification of a specific color interpolation algorithm when it is used for interpolation. We refer to such an algorithm as a *targeted interpolation algorithm*. We model a color interpolation algorithm as a combination of an *architecture* part that entails the algorithmic flow and the *parameter* part that consists of configurable settings. To circumvent the identification, perturbation can be introduced into a parameter part to alter the overall color interpolation algorithm, so that estimated color interpolation coefficients are changed and cannot be recognized by the identification system. As pointed out in Section 6.1, there is a trade-off between the resulting image quality and the manipulation power of identification results. We will examine whether it is possible to reach a good balance between these two factors by wisely selecting the parameters for perturbation.

3.3.1 Perturbing Gradient-based Interpolation

We consider the 5th color interpolation algorithm reviewed in Section 3.2.2 as a targeted interpolation algorithm. This algorithm is based on a gradient-based partitioning of image pixels [2], and its architecture is shown in Fig. 3.2. We consider several options of parameter perturbation that are applicable to this algorithm. First, since the algorithm utilizes bilinear filtering in interpolating the difference between red/green and blue/green channels, one option is to perturb the kernel co-

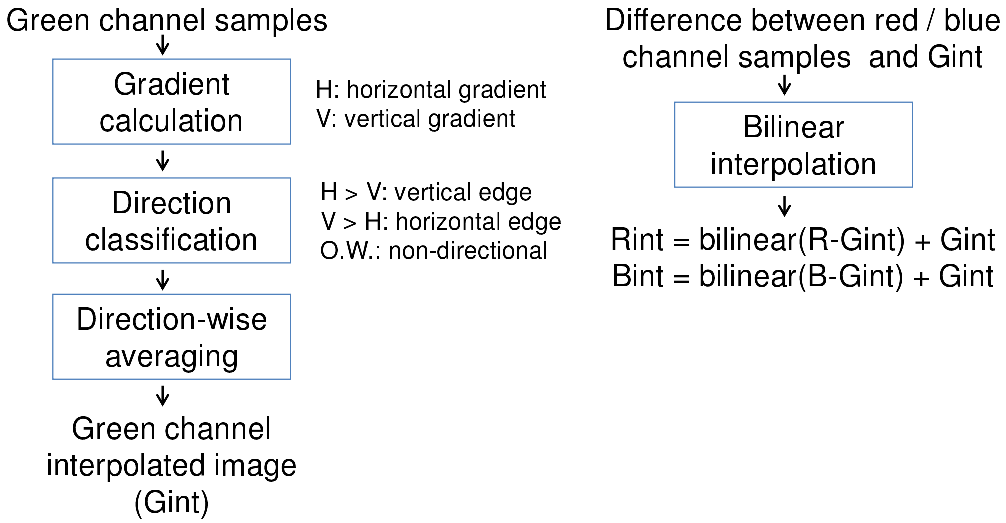


Figure 3.2: Flowchart of Gradient-based color interpolation.

efficients of bilinear filtering. Second, the targeted interpolation algorithm performs pixel averaging in the green channel according to the gradient direction (horizontal, vertical, and non-directional). A second option is hence to perturb the pixel averaging kernels in each direction. Finally, this algorithm takes two parameters, denoted as θ_1 and θ_2, to determine if a pixel falls on a horizontal edge, a vertical edge, or in a non-directional region, so a third option is to perturb the decision boundaries of individual directions. In the summary of these options below, the noise standard deviations are selected so that the trade-offs of different options can be compared more easily:

Option 1: Add Gaussian noise to the bilinear interpolation coefficient matrix. Noise standard deviation $\in \{0.16, 0.24, 0.3\}$. Note that the perturbation has to satisfy constraints on the coefficients' mutual relations. In particular, two coefficients at opposite horizontal/vertical positions, and four coefficients at opposite diagonal positions, must have a fixed sum of 1.

Option 2: Add Gaussian noise to the direction-wise averaging coefficients. Noise standard deviation $\in \{0.1, 0.3, 0.5\}$. Similar to Option 1, a fixed sum constraint must be imposed on the coefficients.

Option 3: Add Gaussian noise to the gradient decision threshold values θ_1 and θ_2. Noise standard deviation $\in \{0.1, 0.15, 0.2\}$. θ_1 and θ_2 must satisfy $\theta_1 + \theta_2 > 0$ so that pixels are assigned into non-overlapping gradient directions.

For comparison, we consider alternative options that do not involve parameter perturbation. For example, in the scenario of technology infringement, if the risk of being caught is high, one option that a pirating company has is to abandon the targeted interpolation algorithm and adopt another algorithm instead. Other alternative options include applying post-processing operations such as compression and filtering after color interpolation in order to conceal the trace of color interpolation. These three more options are summarized below:

Option 4 (i): Replace the gradient-based targeted interpolation algorithm, which is the 5th among those compared in Section 3.2.2, by another interpolation algorithm $i \in \{1, 2, 3, 4, 6, 7, 8\}$.

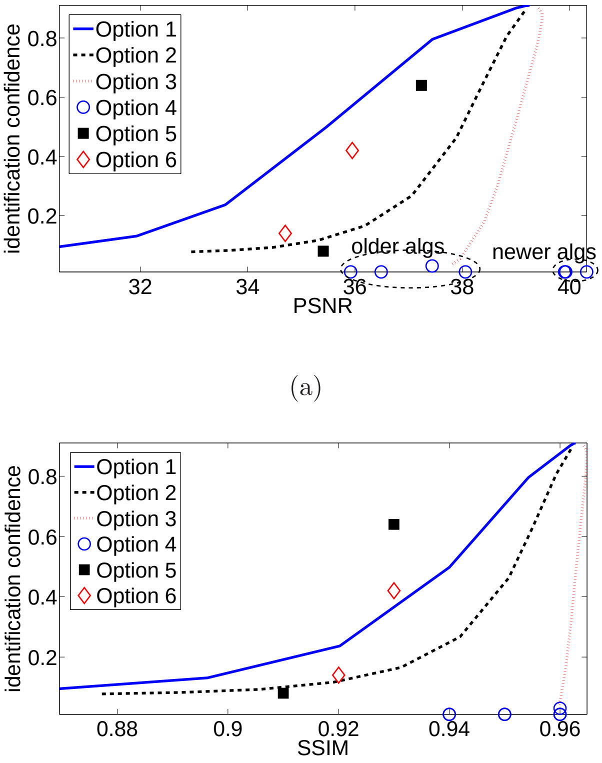
Option 5: JPEG compression after interpolation. Quality factor (QF) $\in \{95, 75\}$.

Option 6 (1): 3×3 median filtering after interpolation; (2): 3×3 average filtering after interpolation.

Comparison of Options: Table 3.1 shows the comparison of various options in terms of image quality and identification confidence. We present multiple image

Table 3.1: Results of countering color interpolation identification for a gradient-based interpolation algorithm. PSNR is measured in dB. “Zipper” stands for the zipper effect ratio; “Conf” stands for the identification confidence.

	Uncompressed				JPEG compressed with QF=95			
	PSNR	SSIM	Zipper	Conf	PSNR	SSIM	Zipper	Conf
Option 1 (1)	38.83	0.96	0.02	0.81	39.50	0.96	0.02	0.81
(2)	38.33	0.96	0.03	0.63	38.99	0.96	0.03	0.70
(3)	37.89	0.95	0.03	0.46	37.16	0.94	0.04	0.66
Option 2 (1)	39.01	0.96	0.02	0.90	39.38	0.96	0.02	0.80
(2)	37.45	0.95	0.03	0.80	37.90	0.96	0.04	0.43
(3)	35.46	0.94	0.05	0.50	36.03	0.94	0.06	0.10
Option 3 (1)	39.02	0.96	0.02	0.53	39.02	0.96	0.03	0.49
(2)	38.66	0.96	0.03	0.30	38.80	0.96	0.03	0.28
(3)	38.41	0.96	0.03	0.18	38.42	0.96	0.04	0.16
Option 4 (1)	35.92	0.94	0.03	0.01	37.33	0.95	0.02	0.03
(2)	36.49	0.95	0.03	0.01	38.04	0.96	0.02	0.01
(3)	37.44	0.96	0.04	0.03	38.08	0.96	0.05	0.04
(4)	38.06	0.94	0.03	0.01	38.23	0.94	0.05	0.01
(6)	39.91	0.96	0.01	0.01	40.20	0.96	0.02	0.02
(7)	39.93	0.95	0.01	0.01	40.03	0.96	0.02	0.01
(8)	40.32	0.96	0.01	0.01	40.54	0.97	0.04	0.01
Option 5 (1)	37.24	0.93	0.02	0.64	38.55	0.95	0.03	0.45
(2)	35.41	0.91	0.03	0.08	37.02	0.94	0.04	0.11
Option 6 (1)	35.95	0.93	0.01	0.42	36.40	0.94	0.02	0.39
(2)	34.70	0.92	0.02	0.14	34.98	0.92	0.02	0.04



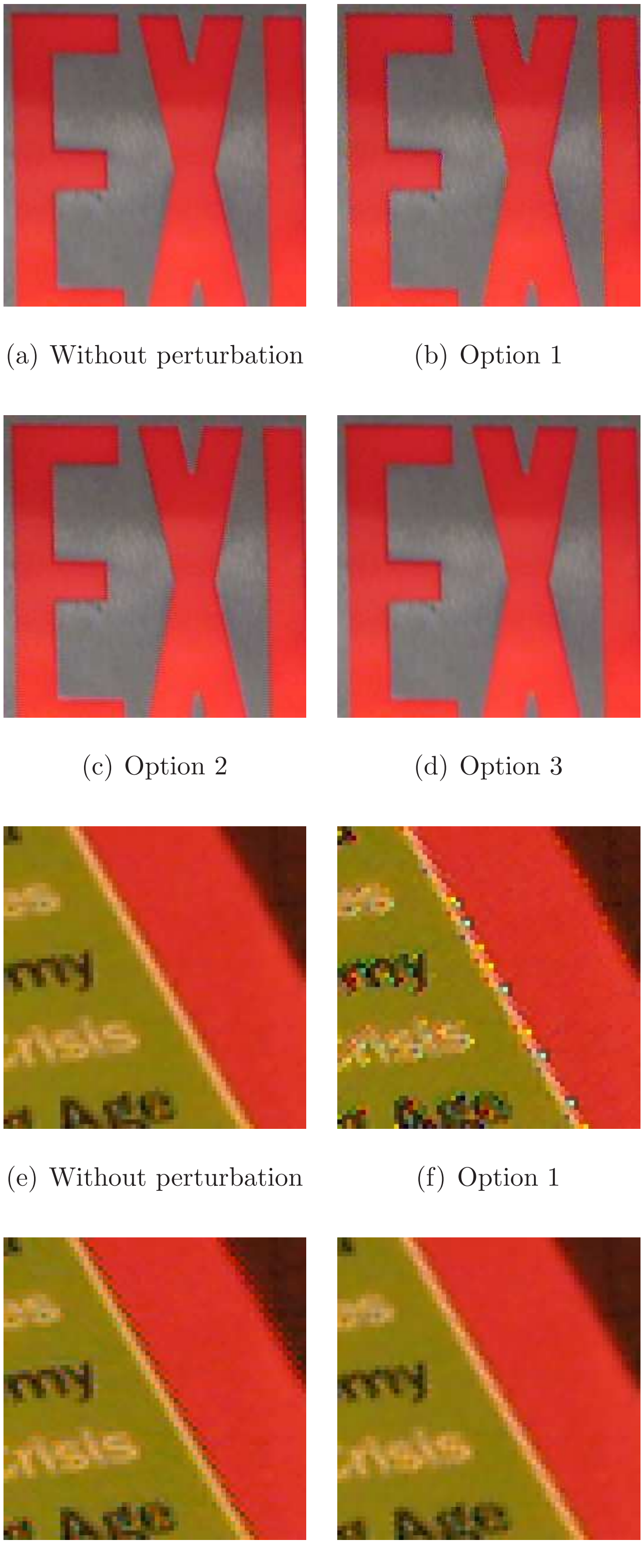
(a)

(b)

Figure 3.3: Visualization of Table 3.1: (a) PSNR versus identification confidence; (b) SSIM versus identification confidence. See Section 3.3.1 for the detailed description.

quality metrics to provide a more comprehensive quality assessment. This table consists of two parts. The left part of columns is the case when there is no post-processing following color interpolation. The right part of columns includes JPEG compression as post-processing. Note that in the right part, the reference image is also compressed. To facilitate the comparison, we also show the relation between 1) PSNR and identification confidence, and 2) SSIM and identification confidence, for varying noise strengths that correspond to the left part of columns.

From Table 3.1 as well as Fig. 3.3, we can first see that parameter perturbation reduces the identification confidence at different costs in terms of image quality.



(a) Without perturbation

(b) Option 1

(c) Option 2

(d) Option 3

(e) Without perturbation

(f) Option 1

(g) Option 2

(h) Option 3

Figure 3.4: Perceptual comparison of images generated by the original interpolation algorithm and Perturbation Options 1, 2, and 3 in Table 3.1 (best viewed on a screen).

Option 2 causes image quality degradation, but the identification confidence is kept relatively high. Note that we have imposed coefficient constraints on Option 1 and 2 to ensure that the perturbed coefficient matrices are still valid; otherwise the unconstrained perturbation would have led to much worse trade-offs between image quality and confidence reduction than the reported values. Compared to Option 1 to 2, Option 3 achieves highest image quality and lowest identification confidence. In particular, Option 3 reduces the identification confidence by 40% with little reduction in image quality (for example, PSNR decreases from 38.66dB to 38.41dB and there is nearly no reduction in other quality metrics). The three options can also be perceptually compared. For the same level of remaining identification confidence (≈ 0.1), we show in Fig. 3.4 two typical images that are generated by the original interpolation algorithm and by each option. It can be easily noticed that in order to effectively reduce the identification confidence, Options 1 and 2 create more artifacts along edges than Option 3, which suggests again that Option 3 achieves a better trade-off between image quality and confidence reduction from an adversary's point of view.

We also compare Option 3 with options that do not involve parameter perturbation. If we replace the gradient-based targeted interpolation algorithm by any other interpolation algorithm as in Option 4, the identification confidence drops to near zero. This is expected since the 8-class pSVM is tailored to differentiate these algorithms. However, for Options 4 (1) to (4) that employ more rudimentary interpolation algorithms, the image quality is inferior to what Option 3 yields, which would be unacceptable as image quality is a crucial criterion in many imaging applications. Option 4 (6) to (8), which replace the gradient-based targeted interpolation algorithm by more sophisticated algorithms, outperform Option 3 in both image quality and identification confidence. This implies that, if a pirating company has more advanced technology, it should utilize such technology and there is no incentive to infringe other companies' technology.

Option 5 and 6 apply post-processing after color interpolation. These options reduce the identification confidence considerably, but none of them produce images with quality comparable to Option 3. Overall, Option 3 that perturbs decision threshold values is simple yet effective for circumventing color interpolation identification with minimal reduction in image quality.

3.3.2 Perturbing Other Interpolation Algorithms

The proposed parameter perturbation methodology is readily applicable to other color interpolation algorithms. In particular, since a majority of interpolation algorithms are direction-adaptive based on local gradients, the options that perturb gradient-related parameters can also be employed. Here we give two more examples in order to further examine the effectiveness of the proposed parameter perturbation technique. We first consider the adaptive color plane algorithm (6th in our list of interpolation algorithms), also known as Hamilton-Adams algorithm [3]. Different from the gradient-based color interpolation algorithm that only involves intra-channel interpolation (*i.e.*, pixels are only interpolated using raw pixels of the same color), the adaptive color plane algorithm also performs inter-channel interpolation (*i.e.*, pixels can be interpolated using raw pixels of different colors). Similar to perturbing the gradient-based algorithm, there are a few options that can be considered. Option 1 and 2 perturb the intra-channel and inter-channel pixel averaging kernels, respectively. Option 3 perturbs the gradient decision threshold values as in the gradient-based interpolation algorithm. The same Options 4 to 6 as in the gradient-based interpolation algorithm are also included for comparison. The results shown in Table 3.2 are consistent with what have been observed in Table 3.1, and we can see that Option 3 that perturbs the gradient decision boundaries is still the most effective choice for circumventing identification while preserving the image quality.

We have also applied parameter perturbation to the LDI-NAT algorithm (our 8th algorithm), which is considered as the state-of-the-art progress in color interpolation [80]. The LDI-NAT algorithm first conducts directional interpolation by assigning relative weights to pixel value estimates along different directions (north, south, east, west), wherein the weights are inversely proportional to local gradient values along respective directions. Then the interpolation results are further enhanced using a nonlocal patch estimation method based on dictionary learning. Compared to the gradient-based or the adaptive color plane algorithm, directional interpolation in the LDI-NAT algorithm does not employ hard partitioning of pixel directions. Therefore, instead of perturbing decision boundaries which are not defined in the LDI-NAT algorithm, we can perturb the weights assigned to respective directions. We consider here for illustration an extreme case of taking as weights the gradient values rather than their reciprocals as in the original LDI-NAT. We

Table 3.2: Results of countering color interpolation identification for the adaptive color plane interpolation algorithm. PSNR is measured in dB. "Zipper" stands for the zipper effect ratio; "Conf" stands for the identification confidence.

	Uncompressed				JPEG compressed with QF=95			
	PSNR	SSIM	Zipper	Conf	PSNR	SSIM	Zipper	Conf
Option 1 (1)	38.04	0.94	0.03	0.94	38.27	0.95	0.04	0.86
(2)	33.12	0.87	0.14	0.90	33.53	0.88	0.16	0.70
(3)	29.64	0.80	0.25	0.75	30.22	0.80	0.27	0.43
Option 2 (1)	38.04	0.94	0.04	0.84	39.25	0.95	0.03	0.69
(2)	37.08	0.93	0.06	0.81	38.85	0.95	0.04	0.66
(3)	36.12	0.92	0.09	0.76	36.66	0.93	0.08	0.45
Option 3 (1)	39.70	0.96	0.01	0.44	39.83	0.96	0.02	0.39
(2)	39.56	0.96	0.01	0.33	39.71	0.96	0.02	0.29
(3)	39.38	0.96	0.01	0.20	39.55	0.96	0.02	0.18
Option 4 (1)	35.92	0.94	0.03	0.01	37.33	0.95	0.02	0.03
(2)	36.49	0.95	0.03	0.01	38.04	0.96	0.02	0.01
(3)	37.44	0.96	0.04	0.03	38.08	0.96	0.05	0.04
(4)	38.06	0.94	0.03	0.01	38.23	0.94	0.05	0.01
(5)	39.24	0.96	0.02	0.01	39.61	0.96	0.02	0.02
(7)	39.93	0.95	0.01	0.01	40.03	0.96	0.02	0.01
(8)	40.32	0.96	0.01	0.01	40.54	0.97	0.04	0.01
Option 5 (1)	37.24	0.93	0.02	0.64	38.55	0.95	0.03	0.45
(2)	35.41	0.91	0.03	0.08	37.02	0.94	0.04	0.11
Option 6 (1)	35.84	0.93	0.01	0.39	36.44	0.94	0.01	0.42
(2)	34.62	0.92	0.02	0.08	35.07	0.93	0.02	0.02

Table 3.3: Results of countering color interpolation identification for the LDI-NAT algorithm. PSNR is measured in dB. “Zipper” stands for the zipper effect ratio; “Conf” stands for the identification confidence.

	Uncompressed				JPEG compressed with QF=95			
	PSNR	SSIM	Zipper	Conf	PSNR	SSIM	Zipper	Conf
Option A	38.50	0.95	0.02	0.50	38.85	0.96	0.03	0.48
Option 5 (1)	37.89	0.94	0.01	0.59	40.37	0.97	0.01	0.58
(2)	35.60	0.91	0.03	0.63	37.17	0.94	0.03	0.52
Option 6 (1)	35.94	0.93	0.01	0.50	36.53	0.94	0.01	0.31
(2)	34.54	0.91	0.02	0.31	34.95	0.93	0.02	0.18

compare this option (denoted by Option A) to JPEG compression (Option 5) and filtering (Option 6) in Table 3.3; note that the same Option 5 and 6 have also been applied to the gradient-based and the adaptive color plane algorithms. We can see that perturbing the gradient-based weights achieves a better trade-off (between image quality and manipulation of identification confidence) than JPEG compression. On the other hand, as it results in a slightly higher identification confidence than filtering, the image quality is substantially higher, too. It can also be observed that, as directional interpolation is only part of LDI-NAT, perturbing its parameters may cause a smaller reduction in the identification confidence.

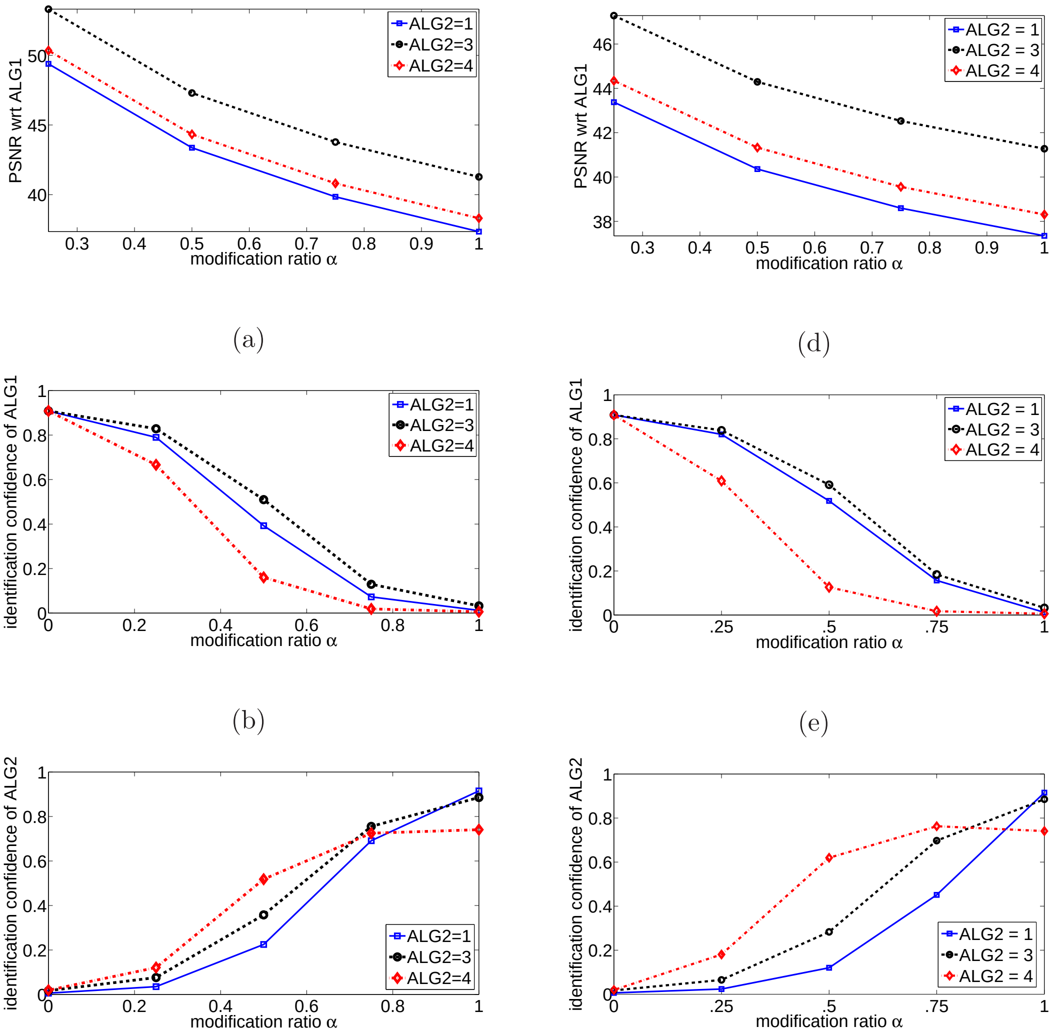
3.4 Misleading Color Interpolation Identification via Algorithm Mixing

So far, we have investigated ways to prevent the color-interpolation-based identification system from identifying a specific interpolation algorithm. We now study how to further mislead the identification system toward a wrong direction, namely, keeping the resulting image visually similar to the original version interpolated by a specific algorithm (referred to as ALG1), while making the identification system believe that the image is interpolated by a different algorithm (referred to as ALG2). This can be considered as a generalized scenario of the one described in Kirchner and Böhme's work [34], wherein ALG2 is the bilinear interpolation. For our study here, the similarity between two images is measured in terms of PSNR, but other metrics such as the SSIM can also be used for similarity measurement.

We examine the fusion of ALG1 and ALG2 per a given *modification ratio* α, $0 \leq \alpha \leq 1$. Specifically, we realize the fusion by mixing pixels generated by ALG1 and ALG2. There are multiple ways to carry out the mixing. One option is to mix pixels interpolated by ALG1 and ALG2 via linear averaging with weights $(1-\alpha)$ and α, respectively. This is is also known as *alpha blending* in the literature of image editing. Alternatively, one can randomly select pixels from ALG1 and ALG2 with ratios $(1-\alpha)$ and α, respectively, which can be seen as non-linear mixing. We examine linear and random mixing methods for the case ALG1=5 and ALG2 $\in \{1, 3, 4\}$ (that is, ALG1 is the 5th algorithm and ALG2 are the 1st, 3rd, and 4th algorithms from Section 3.2.2), while similar results can be observed for other combinations of ALG1 and ALG2 as well. As shown in Fig. 3.5, for both mixing methods, when the modification ratio α increases, the resulting image becomes less similar to the original version by ALG1, the identification confidence of ALG1 decreases, and the identification confidence of ALG2 increases. The exact identification manipulation power at the cost of visual similarity reduction depends on the choice of ALG2. For example, when the modification ratio is 0.5, the choice of ALG2=4 (*i.e.*, the median filter based algorithm) is better at lowering the identification confidence of ALG1 and raising the confidence of ALG2.

On the other hand, these two mixing methods also differ in their trade-offs between visual similarity reduction and identification manipulation. For the illustrative case of ALG1=5 and ALG2=3, Fig. 3.6 shows the relation between the visual similarity to ALG1 and the identification confidence of ALG2. For a given modification ratio α, though these two mixing methods lead to similar identification confidences of ALG2, linear mixing yields a higher PSNR, meaning that the output of linear mixing remains more similar to the output of ALG1.

We also find that algorithm mixing can be employed as an option for circumventing the identification of a specific color interpolation algorithm (namely, the task in Section 3.3). For illustration, we perform algorithm mixing by choosing the gradient-based algorithm as ALG1 and the median filter based algorithm (the 4th in Section 3.2.2) as ALG2. Fig. 3.7 shows the resulting image quality and identification confidence of the targeted interpolation algorithm. Note that if linear mixing is used, the PSNR does not decrease but actually increases when $0 < \alpha < 0.75$. A similar observation has also been reported in [40], and this can be potentially attributed to the independence of interpolation errors between different color interpolation algorithms. For the selected ALG1 and ALG2, both mixing methods achieve better balances between the image quality and the identification confidence as compared to the options considered in Section 3.3.1. For example, for a PSNR value of 38.41dB (the 3rd row associated with Option 3 in Table 3.1), the identi-



(a)

(d)

(b)

(e)

(c)

(f)

Figure 3.5: Algorithm mixing for misleading identification. Left column ((a)-(c)): linear mixing; right column ((d)-(f)): random mixing. (a) and (d): average PSNR with respect to ALG1; (b) and (e): identification confidence of ALG1; (c) and (f): identification confidence of ALG2.

fication confidence yielded by Option 3 is 0.18, but the two mixing methods lead to even lower confidences of 0.09 and 0.01, respectively. Fig. 3.8 shows the average image quality gain in terms of PSNR due to linear mixing. We further examine the

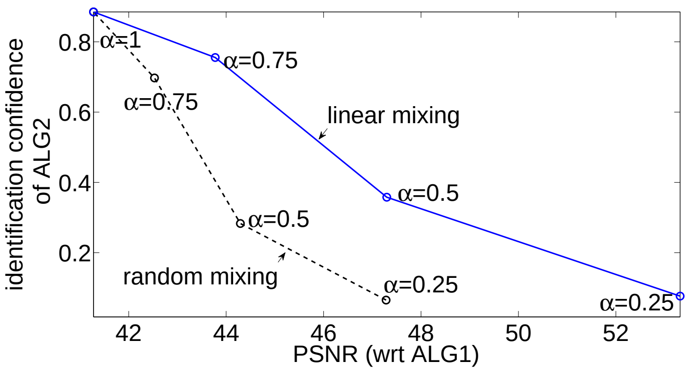


Figure 3.6: PSNR w.r.t. ALG1 versus identification confidence of ALG2. ALG1=5, ALG2=3.

extent of image quality improvement due to linear mixing. Specifically, for each pair of interpolation algorithms, the image quality gain is defined as the non-negative PSNR increase when the two algorithms are linearly mixed with an optimal modification ratio, and the average gain with respect to a given algorithm is obtained by averaging over all pairs that include the given algorithm. As we can see in Fig. 3.8, the median filter based algorithm yields the largest gain (near 0.5dB), suggesting that linearly mixing a targeted algorithm with the median filter based algorithm is a promising option for circumventing identification while preserving (and potentially increasing) the image quality. As a remark, it should be noted that algorithm mixing, especially linear mixing, may require more processing and storage power in the camera since multiple color interpolation algorithms may need to be performed at each pixel location.

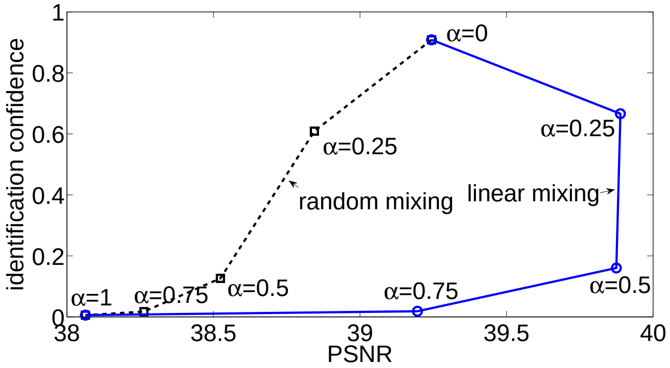


Figure 3.7: Algorithm mixing for circumventing the identification of the gradient-based interpolation algorithm.

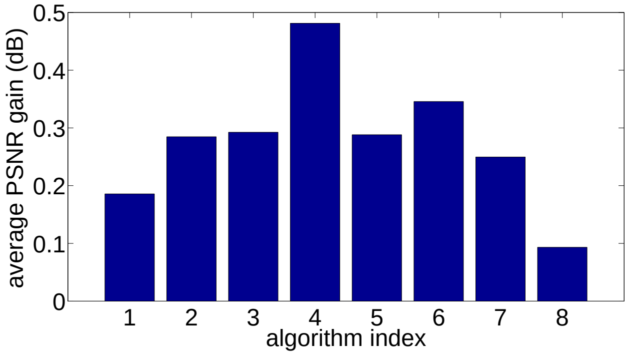


Figure 3.8: Average image quality gain in PSNR due to linear mixing.

3.5 Extensions and Further Discussions

In this section, we provide additional discussions of the proposed anti-forensic techniques. First, we complement the randomized parameter perturbation by formulating and solving an optimization problem that incorporates image quality and identification confidence. We also compare this chapter and a relevant prior work [34]. We then look into the inherent issues and its implications of the state-of-the-art identification system. Finally, we study possible strategies of forensic analysts and adversaries in view of these issues, and characterize their interplay using game-theoretic techniques.

3.5.1 Optimization Problem Formulation of Parameter Perturbation

As an illustrative example, we have applied in Section 3.3 randomized parameter perturbation to conceal the gradient-based color interpolation algorithm, and the performances in terms of the image quality and the identification confidence, are measured by averaging over all the test images. When some images are used for identification, as shown in Fig. 3.9(a), the identification confidence may remain high after the randomized perturbation. In order to ensure identification circumvention for individual images, note that the identification is usually performed by an automated detector, and thus it is sufficient and necessary to make the identification confidence fall below a threshold that has been set in the automated detector. Toward this end, we formulate parameter perturbation as the following optimization problem:

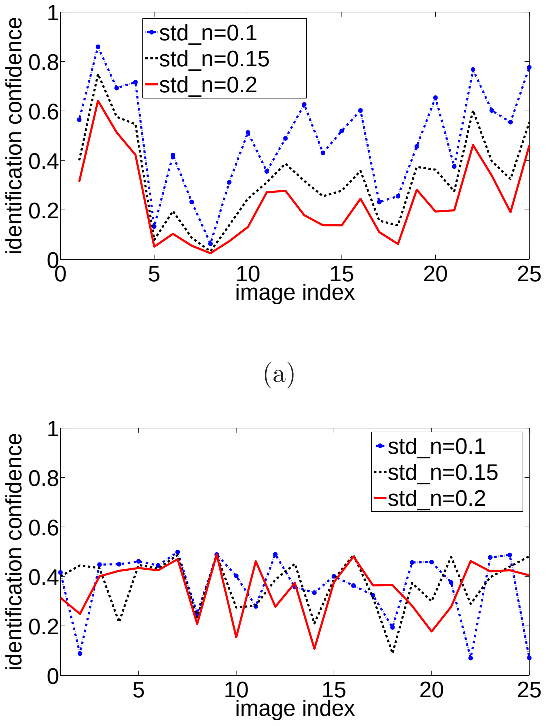
$$\max_{\theta_1,\theta_2} Q(I_p), \quad \text{subject to } C(I_p) \leq C_t,$$

where I_p is the perturbed image, $Q(\cdot)$ is a quality metric of an image, $C(\cdot)$ is the identification confidence with respect to a targeted interpolation algorithm, and C_t is a preset threshold. As the full-color reference image is not available during color interpolation, we adopt the image that is interpolated by the original gradient-based interpolation algorithm an approximate reference image in the optimization. The PSNR with respect to this reference image is taken as the quality metric $Q(\cdot)$, and $C(\cdot)$ comes from the identification confidence of the gradient-based algorithm reported by the 8-class pSVM.

Since it is not always feasible to represent $Q(I_p)$ and $C(I_p)$ in a closed form, solving for the perturbation parameters θ_1 and θ_2 is a challenging optimization task. In this chapter, we take a Monte-Carlo approach that applies Option 3 in Section 3.3.1 multiple times to perturb the image, and keep the result that satisfies the constraint on $C(I_p)$ with highest $Q(I_p)$. Compared to randomized perturbation, this solution is guided explicitly by the image quality and the identification confidence. We compare the results of Option 3 and the guided perturbation when $C_t = 0.5$ for three different noise strengths. Their average PSNR values are roughly equal. The identification confidences are shown in Fig. 3.9. It can be seen that the proposed approach suppresses the identification confidence for individual images while maintaining a high image quality; the results also suggest that the approximation of the reference image by the image interpolated using the gradient-based algorithm is effective.

3.5.2 Comparison with Kirchner and Böhme [34]

As reviewed in Section 6.1, the work by Kirchner and Böhme [34] is a related prior work that studies anti-forensic techniques for color interpolation identification. Despite the similar goal, the approaches adopted in [34] and the present chapter differ substantially. Kirchner and Böhme's work tries to synthesize a linear dependency among pixels in an image while minimizing the overall distortion. The



(a)

(b)

Figure 3.9: Identification confidences as a result of randomized parameter perturbation (a) and the guided parameter perturbation (b). Identification confidence in (b) ≤ 0.5.

authors proposed to search for a pre-filter that estimates raw samples acquired by the camera sensor array and applies the bilinear interpolation kernel to the estimated raw samples to reconstruct the entire image that satisfies the linear dependency. This approach can be viewed as altering the raw samples to counter the identification of color interpolation. In contrast, our proposed approaches leave the raw samples unchanged, but alter the color interpolation algorithms so that the output image either deviates from a target color interpolation algorithm or moves toward the algorithm. It can be viewed that Kirchner and Böhme's method alters the color interpolation *after* the creation of an image, while our techniques alter the color interpolation *during* the creation of an image. Also notice that in Kirchner and Böhme's work, even for the case of bilinear interpolation, searching for the pre-filter (or equivalently, the virtual raw samples) is already computationally challenging, and it becomes even more difficult to generalize this method to more sophisticated color interpolation. In comparison, our techniques are less complex and exhibit a promising generalization capability. It will be an interesting future work to explore whether Kirchner and Böhme's work and our approaches can be properly fused for improved anti-forensic capability.

3.5.3 Reflections on Resilience of Color Interpolation Identification

As motivated in Section 6.1, a fundamental reason for studying anti-forensic operations against color interpolation identification is to understand the resilience of identification schemes in an adversarial environment against intentional manipulations of identification results. As demonstrated in the chapter, properly configured parameter perturbation and algorithm mixing can circumvent and mislead the identification system while preserving image quality.

We have observed that by perturbing the decision boundaries of gradient directions, the identification confidence can be reduced with minimal reduction in image quality. The rationale of such effectiveness can be understood as follows. In order to capture the nature of direction adaptation in prevailing color interpolation algorithms (for example, the gradient-based, adaptive color plane, and

LDI-NAT algorithms considered in this chapter), today's color interpolation identification schemes [10, 67] are primarily based on direction classification of pixels and least-squares estimation of interpolation coefficients for each class. By perturbing the decision boundaries in color interpolation, we are essentially changing the ways some pixels are interpolated, and this directly makes the estimated color interpolation coefficients deviate from the typical values learnt from the original color interpolation algorithm, reducing the identification capability. In the meantime, pixels whose interpolation are more likely to be changed are those near the decision boundaries. These pixels are not coupled tightly with respective direction classes in the interpolation algorithm, and none of the classes is likely to interpolate these pixels particularly well. As such, the image quality does not seriously degrade when these pixels are interpolated by the methods associated with other direction classes. On the other hand, our investigation of algorithm mixing, especially linear mixing, suggests the possibility of manipulating identification results while potentially increasing the image quality. This can be attributed to the independence of interpolation errors caused by individual interpolation algorithms, and one could effectively counter the identification by properly selecting the modification ratio, given the validity of error independence. With our work raising the awareness of these inherent and common issues of color interpolation identification, forensic researchers could improve identification techniques accordingly to combat anti-forensics.

3.5.4 Color Interpolation Identification Game

As discussed in Section 3.5.3, because color interpolation identification based on directional classification is sensitive to pixels near the decision boundaries of gradient directions, perturbing the decision boundaries can reduce identification confidence while preserving the image quality. In order to address such vulnerability, a forensic analyst can ignore or treat with lower weights those pixels near boundaries when estimating the color interpolation coefficients. This may make the identification system more resilient in the presence of the proposed anti-forensic operation, but may reduce the estimation accuracy in the absence of anti-forensics. On the other hand, if the adversary is aware of the forensic analyst's countermeasure, he/she may choose to perform a stronger anti-forensic operation that affects more pixels, at a cost of more severe image quality degradation. We can see that there is a dynamic interaction between the forensic analyst and the adversary, and both the forensic analyst and the adversary's actions will depend on each other's action. It is of interest to understand what actions will be eventually taken, and what outcome such actions will lead to. It has been shown in recent years that game theory [49] is a powerful tool for studying strategic decision making, and we formulate a color interpolation identification game to address the questions raised above. Without loss of generality, we will focus on the scenario where the forensic task is to develop a color interpolation based detector that distinguishes the gradient-based color interpolation algorithm among others that are listed in Section 3.2.2.

Denote the forensic analyst and the adversary by Player FA and Player AD, respectively. In the interaction between the two players, Player FA's strategy selects the pixels that will be used for estimating the color interpolation coefficients. More specifically, for Player FA, we define the *typicality* for pixels associated individual direction regions as follows:

$$T_{x,y} = \begin{cases} V_{x,y} - H_{x,y}, & \text{if}(x,y) \in R_1; \\ H_{x,y} - V_{x,y}, & \text{if}(x,y) \in R_2; \\ A_{x,y} - D_{x,y}, & \text{if}(x,y) \in R_3; \\ D_{x,y} - A_{x,y}, & \text{if}(x,y) \in R_4; \\ (V_{x,y} + H_{x,y} + A_{x,y} + D_{x,y})^{-1}, & \text{if}(x,y) \in R_5, \end{cases}$$

where $V_{x,y}$, $H_{x,y}$, $A_{x,y}$, and $D_{x,y}$ are defined as in Section 3.2.1. A high typicality means that the pixel is a typical sample of its associated direction region and is far from the decision boundary. Player FA's strategy selects pixels by sorting all pixels according to their typicality and picking α_T% of pixels with highest typicality, where $1 \leq \alpha_l \leq \alpha_T \leq 100$. The lower limit α_l is imposed to ensure that there are enough pixels and the color interpolation coefficient estimation is not ill-conditioned. On the other hand, Player AD's strategy selects the noise strength, denoted by S_n, in the Option 3 described in Section 3.3.1.

For a given pair of strategies (α_T, S_n), the utility that Player FA will maximize is the identification confidence $C(\alpha_T, S_n)$, *i.e.*,

$$U_{\text{FA}}(\alpha_T, S_n) = C(\alpha_T, S_n).$$

In contrast, Player AD will minimize the identification confidence while taking additional care of the image quality. The exact utility function associated with Player AD depends on the exact problem settings. For example, if Player AD can only minimize the identification confidence subject to a specified constraint Q_t on the image quality $Q(\alpha_T, S_n)$, then the utility function can be written as

$$U_{\text{AD}}(\alpha_T, S_n) = -C(\alpha_T, S_n) \times \mathbb{1}\left(Q(\alpha_T, S_n) \geq Q_t\right).$$

where $\mathbb{1}(\cdot)$ is the indicator function. Since $Q(\alpha_T, S_n)$ is independent of α_T and is a decreasing function of the applied noise strength S_n, this utility function can be rewritten in terms of a noise strength constraint S_t:

$$U_{\text{AD}}(\alpha_T, S_n) = -C(\alpha_T, S_n) \times \mathbb{1}\left(S_n \leq S_t\right). \tag{3.1}$$

A key concept in game theory is the Nash equilibrium, which is a particular selection of both players' strategies with the property that any unilateral strategy change by a player cannot increase the player's utility. As such, the Nash equilibrium stands for a stable pair of strategies that both players would have the incentives to adopt. For the utility function in (3.1), since the indicator function essentially limits the range of S_n that leads to a non-zero utility, we can ignore the indicator function by constraining Player AD's possible strategy: $S_n \in [0, S_t]$. As a result, the game is simplified as a zero-sum game, whose Nash equilibrium can be readily found as the minimax solution:

$$(\alpha_T^*, S_n^*) = \arg \max_{\alpha_l \leq \alpha_T \leq 100} \min_{0 \leq S_n \leq S_t} C(\alpha_T, S_n).$$

For the range of $\alpha_l \leq \alpha_T \leq 100$ where $\alpha_l = 30$ and $0 \leq S_n \leq 0.2$, we show in Fig. 3.10 the numerical evaluation results of $C(\alpha_T, S_n)$. In this figure, each curve represents the relation between $C(\alpha_T, S_n)$ and α_T for a fixed S_n; the step size of S_n between adjacent curves is 0.01. On one hand, as we have discussed in Section 3.3.1, increasing S_n always reduces the identification confidence. Therefore, under our setting, Player AD has the incentive to increase S_n as long as it does not exceed S_t. On the other hand, the way α_T affects the identification confidence is a function of S_n. When S_n is small (e.g., $S_n = 0$), the identification confidence remains unchanged if α_T is large and then decreases as α_T decreases. This implies that 1) pixels that are closest to the decision boundaries are not useful for estimating the color interpolation coefficients and therefore can be ignored during the estimation; 2) pixels far from the decision boundaries (*i.e.*, typical pixels) should be included in the estimation otherwise the identification confidence will decrease. In contrast, when S_n is large (e.g., $S_n = 0.2$), the identification confidence increases as α_T decreases, meaning that more pixels near the decision boundaries should be ignored in the estimation as they are highly likely to be perturbed. For a moderate value of S_n (e.g., $S_n = 0.1$), the identification confidence increases as α_T decreases for larger α_T, and the identification confidence decreases as α_T decreases for smaller α_T. As a general principle, it can be seen that there is an optimal value of α_T that should be taken by Player FA, which also depends on S_n taken by Player AD. From Fig. 3.10, it is clear that the Nash equilibrium can be achieved by letting Player AD take the maximum allowable S_n and then letting Player FA take the optimal α_T accordingly. At the Nash Equilibrium, notice that Player AD can suppress the identification confidence substantially if a lower image quality is allowed; this is in line with the fact that perturbing the decision boundaries is a very effective

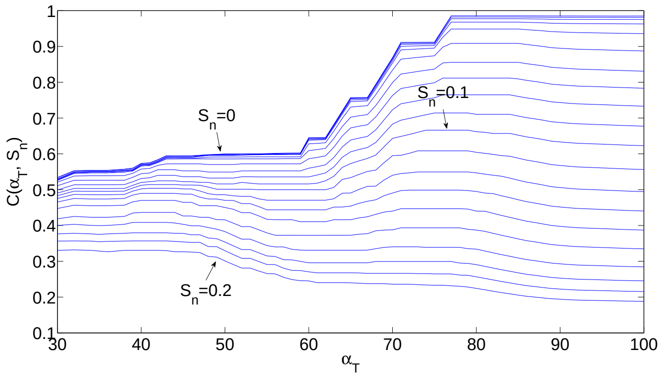


Figure 3.10: Identification confidence as a function of the typicality percentage threshold α_T and the noise strength S_n.

anti-forensic technique. Nevertheless, a proper choice of α_T can still increase the identification confidence. For example, when $S_t = 0.1$, choosing $\alpha_T \approx 76$ can increase the identification confidence by 4% as compared to $\alpha_T = 100$, and when $S_t = 0.2$, choosing $\alpha_T \leq 42$ increases the identification confidence by 14%. As a final remark, note that the proposed color interpolation identification game can be adapted to other settings if the utility functions are redefined accordingly, such as in [62] where the identification performance and the resulting image quality are fused in the adversary's utility function in an additive manner.

3.6 Chapter Summary

Identification of color interpolation has been shown to be a promising approach to assisting forensic analysis regarding imaging devices and content. However, in order to ensure the trustworthiness of forensic identification especially in an adversarial environment, it is necessary to understand how color interpolation identification performs against anti-forensic operations that manipulates identification results.

In this chapter, we have proposed two techniques for countering color interpolation identification. For the technique of parameter perturbation, we have examined options that achieve different trade-offs between two important factors, the image quality and the reduction in identification confidence. We show that perturbing the decision threshold values for pixel classification is a simple yet effective option for circumventing the identification. For the technique of algorithm mixing that fuses results from multiple algorithms, we have quantitatively compared different mixing settings and shown that it is feasible to further mislead the identification system while preserving the image quality.

To complement the randomized nature of the parameter perturbation technique, we have formulated it as an optimization problem and proposed a Monte-Carlo type of approach that maximizes individual image quality with the identification confidence kept low. We have also compared our proposed anti-forensics with the most relevant work [34], and found that our approach has the advantages of lower complexity and better generalization capability. Based on the analysis presented in this chapter, we have shed light on the inherent issues of the current identification system that has performed well. Such an insight has been further formulated as a game of color interpolation identification wherein optimal strategies that the forensic analyst and the adversary can take have been studied. We envision that the proposed methodology can be applied to examine other imaging processes, and forensic researchers can exploit the understanding of anti-forensics as guidelines to design more resilient techniques for digital imaging identification.

CHAPTER 4

Electrical Network based Time Stamping against Anti-Forensics

4.1 Chapter Introduction

The recent decade has witnessed a huge amount of multimedia data, in the form of audio, image, and video, created by various digital recording devices. Once a multimedia document containing important information is created, it can be easily distributed through network and social media infrastructure and make rapid and broad social impacts. However, the digital nature of multimedia data makes it vulnerable to digital forgeries. For example, many digital editing software packages can be used to cut a clip from one audio/video file and insert into another, or to modify the creation date/time in the metadata field. In view of the feasibility of digital forgeries, reliable use of multimedia data requires forensic authentication mechanisms that can identify data origin and detect content tampering.

One emerging direction of digital recording authentication is to exploit a time stamp originated from the electrical network. This time stamp, referred to as the electrical network frequency (ENF) signal, is based on the fluctuation of the supply frequency of a power grid. The nominal value of the ENF is 60Hz in the Americas, Taiwan, Saudi Arabia and Philippines, and is 50Hz in other regions except Japan, which adopts both frequencies. It has been found that digital devices such as audio recorders, CCTV recorders, and camcorders that are plugged into the power systems or are near power sources may pick up the ENF signal due to the interference from electromagnetic fields created by power sources [27]. An important property about the ENF signal is that its frequency is fluctuating around the nominal value because of varying loads on the power grid. For example, in the United States, the ENF usually varies between 59.9Hz and 60.1Hz. It has also been shown that the fluctuations measured at the same time but at two different locations under the same power grid follow basically a similar trend [27].

The fluctuation of the ENF has been successfully exploited to authenticate digital recordings [25, 27, 59, 60]. In [27, 60], it is demonstrated that the ENF signal is captured in audio recordings and exhibits a high correlation with the ENF signal measured from the power mains supply at the same time. As such, the ENF signal can be used to indicate the creation time of an audio recording provided that a database of ground-truth ENF signals from the power grid is accessible. An alternative technique in [59] detects the phase discontinuity of the ENF signal, whose presence suggests where tampering has taken place. Most recently, the work in [25] validated for the first time the presence of the ENF signal in visual recordings. Optical sensors and video cameras are used to demonstrate that the ENF signal can be captured from fluorescent lighting and further picked up by video cameras in an indoor environment. This finding suggests that the same ENF-based time stamp can be used to authenticate visual data as well. Furthermore, forensic binding of visual and audio tracks can be performed to verify their temporal synchronization [25].

The promising potential of ENF analysis in forensic investigations is based on the premise that the ENF signal is present in an audio or video signal in an unaltered manner. This premise ensures that once the ENF signal is successfully extracted, it can be used as a truth-telling evidence to verify the recording time, location, and data integrity. However, similar to many other security and forensics tasks, there exist adversaries who have the incentives to perform *anti-forensic operations* to counteract forensic investigations [18,35]. In order to establish ENF-based analysis as a credible technique, it is of paramount importance to understand its robustness against anti-forensic operations, namely, whether the ENF signal can be compromised, and to what extent. Further, forensic analysts should understand and address identified vulnerabilities in ENF analysis, and take into consideration possible improvements that an adversary may make. Anti-forensic operations can be grouped into physical means and digital processing. The current chapter is a comprehensive development based on the preliminary work in [14], which, to the best of our knowledge, is the first work that considers digital-domain anti-forensics of ENF-based analysis. We investigate anti-forensic operations that are based on signal processing techniques, and then devise detection methods targeting these operations. In response to the detection methods, concealment methods are also investigated in this chapter, for which various trade-offs are discussed. More fundamentally, we develop a comprehensive understanding of the interplay between the forensic analyst and the adversary, from an evolutionary perspective and a game-theoretic perspective. These perspectives are then applied to study representative scenarios and the corresponding optimal strategies are also developed.

The rest of this chapter is organized as follows. Section 4.2 reviews the mechanism of ENF signal extraction and matching. Section 4.3 investigates ways to remove an ENF signal present in a host signal and embed an alien ENF signal into the host signal. Section 4.4 presents the conditions for anti-forensics detection, which motivate a few concrete methods for anti-forensics detection. In response to the detection, Section 4.5 studies concealment techniques, and discusses various trade-offs. In view of the dynamic nature of the anti-forensics and the countermeasures, Section 4.6 provides an evolutionary perspective and a game-theoretic perspective to encompass a wide range of actions and interactions available to a forensic analyst and an adversary. Representative scenarios are quantitatively studied and optimal strategies are derived. Section 4.7 summarizes this chapter.

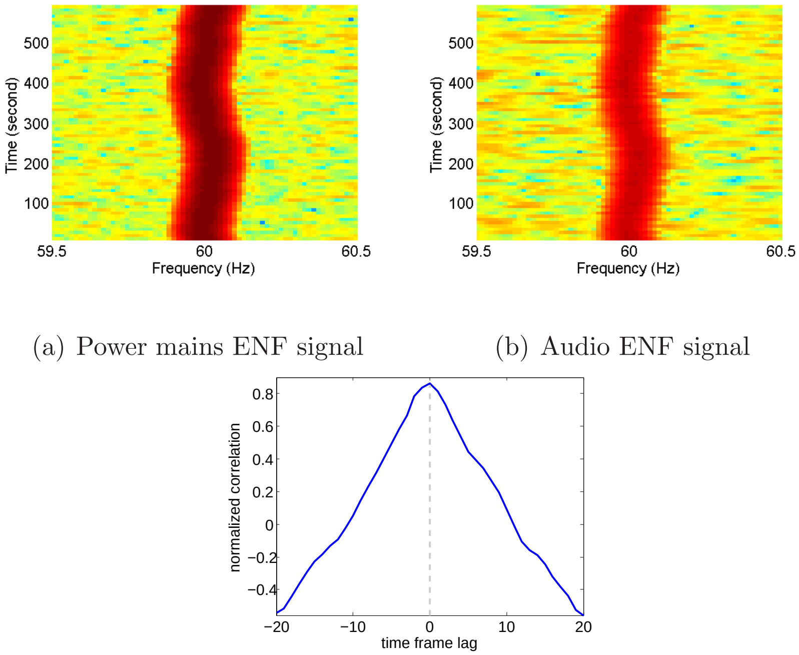
4.2 ENF Signal Extraction and Matching

In this section, we briefly describe our procedure for extracting the ENF fluctuations from a given signal. Two types of signals are considered in this chapter for ENF signal extraction and matching. The first is the audio signal that contains speech recordings mixed with music and sporadic sound activities. All audio signals used in this chapter have been sampled at 8000Hz with 16-bit quantization precision and a length of 10 minutes. The 10-minute duration ensures that the audio signal as well as the ENF fluctuations are sufficiently long for reliable matching based on the state of the art. Any anti-forensic operations to be investigated in this chapter are also assumed to be performed on such audio signals. The second type of signal is the power mains signal that is recorded directly from a power source with a voltage divider device. This type of signal is used as ground truth for matching.

Our ENF signal extraction basically follows the procedure described in [25]. The recorded signal (either an audio or power mains signal) is first down-sampled to 500Hz to reduce the complexity of subsequent filtering and frequency estimation. A filtering process can then be carried out to only retain the signal component that carries the ENF. The dominant instantaneous frequency in the recorded signal is then estimated to measure the fluctuations in ENF as a function of time using the spectrogram based weighted energy method as in [25]. To obtain the spectrogram of the ENF signal, we divide the signal into overlapping frames of 16 seconds each with an overlap factor of 50%. A Fast Fourier Transform (FFT) of 8192 points is carried out for each frame. After obtaining the spectrogram, we calculate the weighted average frequency in each time bin of the spectrogram by weighing frequency bins around the nominal values of the ENF with the energy present in the corresponding frequency. For the estimated frequency fluctuations in ENF signals from the audio and power mains recordings, we calculate their normalized correlation for different values of frame lag. The range of the normalized correlation value is between -1 and $+1$. As an example, Fig. 4.1(a) and 4.1(b) show the spectrograms around the nominal ENF value of 60Hz of a power mains signal and an audio signal that were recorded at the same time. Their normalized correlation values as a function the frame lag is plotted in Fig. 4.1(c). We can see that they exhibit consistent fluctuations, which is confirmed by the peak normalized correlation value of 0.86 in Fig. 4.1(c) when the two recordings are synchronized.

4.3 Anti-Forensic Operations against ENF Analysis

In this section, we investigate anti-forensic operations that can counteract ENF analysis. The general purpose of anti-forensic operations is to alter a host signal so that the traces left in the host signal that pertain to specific forensic investigations are removed or changed. While plausible anti-forensic operations and countermeasures are domain-specific and may seem ad-hoc at times, exploring these operations and countermeasures is necessary for identifying the available operations of both the forensic analyst and the adversary. In many anti-forensic tasks against information protection, the adversary has to preserve the quality of the host signal, otherwise the quality degradation in itself will indicate the use of anti-forensics and



(a) Power mains ENF signal

(b) Audio ENF signal

(c) Normalized correlation

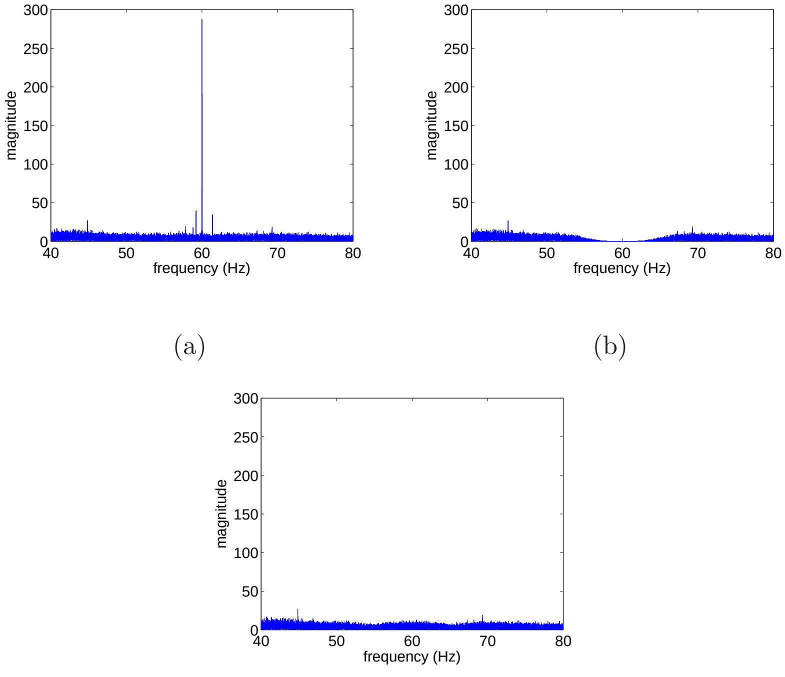
Figure 4.1: (a) Spectrogram of a power mains signal around the nominal ENF value of 60Hz; (b) spectrogram of an audio signal; (c) normalized correlation between the two extracted ENF signals as a function of their relative frame lag.

the host signal will be rejected to be forensic evidence. In our problem, the ENF signal is restricted around narrow neighborhoods of known frequency locations. As such, the ENF signal is less likely to be tightly coupled with the main body of the host signal, making it possible for an adversary to manipulate the ENF signal while trying to preserve the perceptual quality of the host signal. In this section, we explore two different levels of anti-forensics, starting with the removal of the ENF signal and further considering the embedding of an alien ENF signal.

4.3.1 ENF Signal Removal by a Bandstop Filter

The first anti-forensic operation that we consider is to remove the ENF signal present in a host signal. Since the ENF signal in nature is restricted in a small frequency region (a.k.a. *narrowband* hereafter), it is reasonable for an adversary to apply a bandstop filter to remove the ENF signal. Bandstop filtering (a.k.a. notch filtering) is a well-studied subject in digital signal processing [54]. A number of design methodologies, e.g., equiripple filter or Kaiser window filter designs, have been proposed and implemented in popular software packages such as MATLAB. To perform bandstop filtering, an adversary selects two main parameters, the stopband bandwidth and the transition bandwidth. The stopband bandwidth controls the frequency range wherein the signal is attenuated to the minimum magnitude level. For the task of ENF signal removal, the choice of stopband bandwidth depends on the actual range of ENF variation, and ENF signals of wider variations may be removed using wider stopbands. The second parameter, the transition bandwidth, is the range wherein the signal attenuation varies from maximum to minimum. It has an impact on the filter length and computational complexity; a sharper bandwidth implies a longer filter and more time required to compute the filter output. Since accurate ENF matching requires ENF signals of sufficiently long durations, it is reasonable to assume that audio signals used for anti-forensic operations are also sufficiently long. Therefore, if the adversary can afford the computational cost, he/she has enough signal samples to carry out a bandstop filtering with a reasonably small transition bandwidth. As an example, when the sampling frequency is 8000Hz which is common for voice signals, we set the stopband bandwidth as ±1Hz, and the transition bandwidth as 8Hz. If the equiripple linear-phase design is adopted, the filter has a length of 3627 samples, which corresponds to a duration of about half a second.

To illustrate the effect of bandstop filtering, we show in Fig. 4.2(a) a typical Fourier analysis result on a 10-minute audio recording. There is a salient peak located at 60Hz, which signifies the existence of the ENF signal. The effect of bandstop filtering for the same audio recording is shown in Fig. 4.2(b), wherein the peak at 60Hz disappears, suggesting that the ENF signal has been removed. The removal is further justified by comparing the normalized correlation between the ENF signal extracted from power mains ground truth and the ENF signal extracted from the audio recording. We notice that the normalized correlation reduces from 0.86 to -0.10 due to bandstop filtering, suggesting that the ENF signal has been effectively removed. Furthermore, our subjective tests do not find perceptual audio quality loss, meaning that the ENF signal removal preserves the main utility of the



(a)

(b)

(c)

Figure 4.2: (a) The FFT magnitude of an authentic audio clip; (b) the result of bandstop filtering; (c) the result of bandstop filtering followed by noise filling-in.

host signal.

Although bandstop filtering can remove the ENF signal, a notch of very low magnitude around the 60Hz frequency can be noticed in Fig. 4.2(b). The notch is a strong evidence that suggests the use of bandstop filtering, making the resulting audio recording no longer trustworthy and hence anti-forensics essentially fails. To erase such traces, an option is to “fill in” the frequency region that has been suppressed by bandstop filtering. We design a bandpass filter with passband bandwidth ±1Hz and transition bandwidth 8Hz and pass a white noise signal through the filter to obtain a narrowband signal that is then added to the bandstopped audio recording. The noise power is selected so that the resulting narrowband magnitude equals

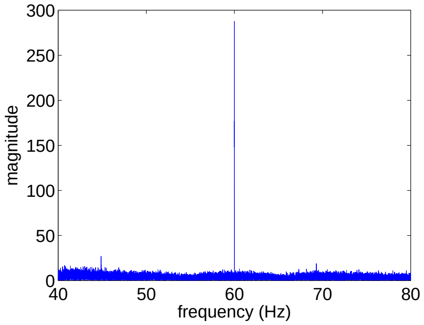


Figure 4.3: ENF embedding result with peak magnitude matched (see Fig. 4.2(a) for comparison).

the average magnitude of neighboring narrowbands, as shown in Fig. 4.2(c). Since the narrowband now appears smooth and there is no peak at 60Hz, it becomes more difficult for the forensic analyst to determine if there was a measurable ENF signal present at 60Hz.

4.3.2 Embedding Phony ENF Signals

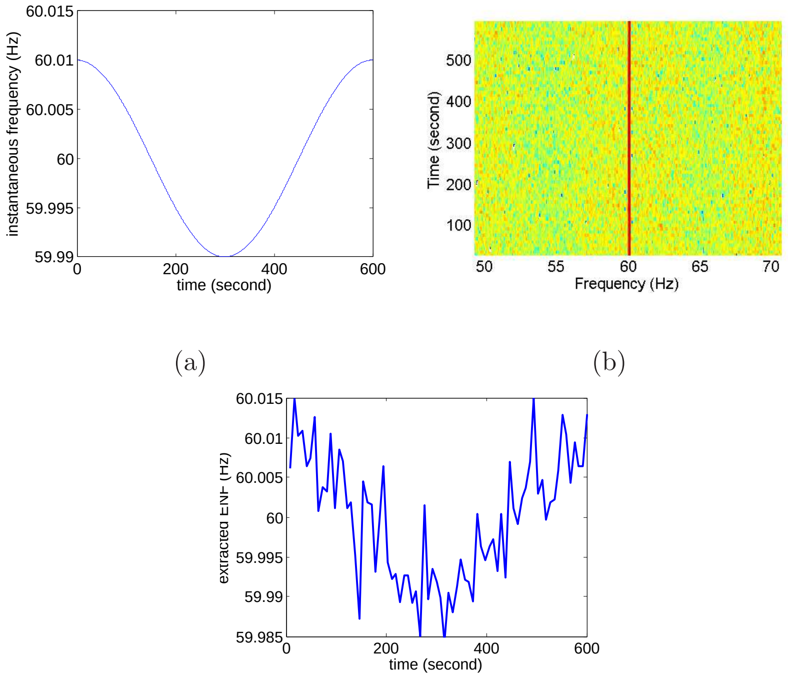
In addition to removing the ENF signal so that the creation time of an audio recording is no longer available, an adversary may further embed a fake ENF signal into a host signal so that ENF analysis conducted over the forged audio signal leads to a wrong estimate for the recording time. This can be done by modulating a carrier sinusoidal signal of a nominal frequency using a given sequence of instantaneous frequencies. In mathematical terms, the carrier signal can be written as $c(t) = M\cos(2\pi f_c t)$, where the magnitude M is a constant to be determined. The modulation is given by

$$e(t) = M\cos\left(2\pi\int_0^t f_m(\tau)d\tau\right), \tag{4.1}$$

which is the standard form of Frequency Modulation (FM) synthesis [29]. Indeed, the instantaneous frequency of (4.1) is given by $\frac{d}{dt}\frac{1}{2\pi}\left(2\pi\int_0^t f_m(\tau)d\tau\right) = f_m(t)$. Next, we discuss how to embed a modulated signal into a host signal. As in Section 4.3.1, we first apply a bandstop filter on the host signal and then fill in bandpassed noise whose magnitude is matched to neighborhood regions. The magnitude M in (4.1) is chosen so that the peak FFT magnitude at the nominal frequency remains the same after the anti-forensic operation, as shown in Fig. 4.3. This can be achieved using a binary search procedure: starting with an arbitrary guess of M, each iteration compares the resulting peak FFT magnitude to the targeted value and increases/decreases M accordingly.

We consider two possible types of synthetic ENF signals. If there is no real ENF signal from another time or another power grid available for embedding, one can embed a purely artificial signal such as the sinusoidal variation as shown in Fig. 4.4(a). The resulting spectrogram has a strong component around 60Hz as shown in Fig. 4.4(b), and the ENF signal extracted from the forged audio signal is shown in Fig. 4.4(c), which is a noisy version of Fig. 4.4(a) since the embedded signal has been mixed into the narrowband. On the other hand, if a real ENF signal originated from a different time or from another power grid is available, then such an ENF signal can also be embedded into the host signal to mislead forensic analysis. Fig. 4.5 shows a power mains ground truth ENF signal, and the corresponding extracted ENF. We can see that the embedded ENF can also be extracted in a more noisy form.

The proposed embedding above is based on the FM synthesis. Alternatively,



(a)

(b)

(c)

Figure 4.4: (a) A purely sinusoidal sequence of instantaneous frequencies to be embedded as the ENF signal; (b) the spectrogram around 60Hz where a strong component is present due to the embedding of (a); (c) the corresponding extracted ENF signal.

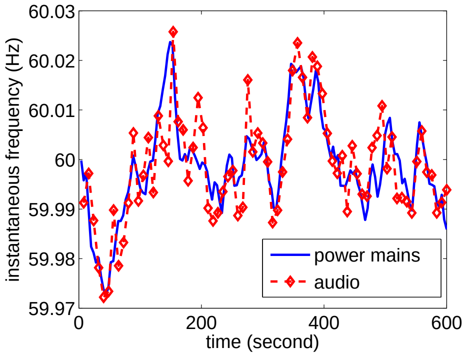
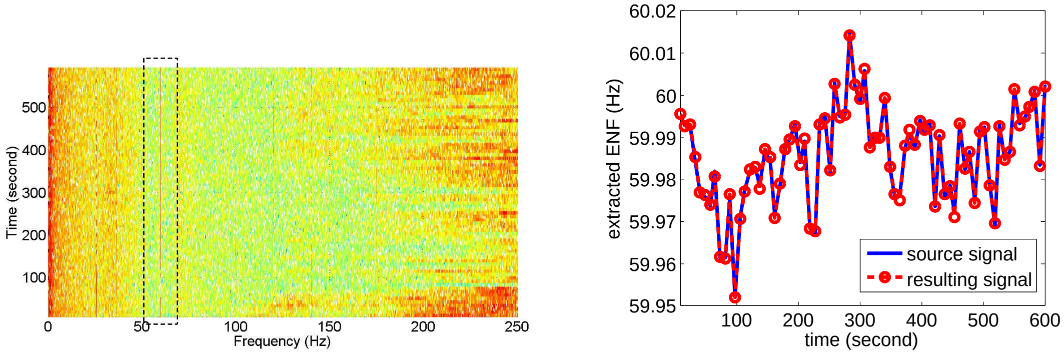


Figure 4.5: Ground-truth ENF signal measured from the power mains (in blue) and the corresponding extracted ENF signal (in red).

one can perform a "transplantation" operation to duplicate the ENF signal from one signal into another signal. Specifically, to embed an ENF signal present in a source audio signal into a host signal, we perform bandpass and bandstop filtering upon the source and the host signal, respectively, and then add the bandpassed output of the source signal into the bandstopped output of the host signal. In Fig. 4.6(a), we show the spectrogram of a transplantation result in which the 60Hz narrowband has been replaced. The extracted ENF signals from the source signal and the resulting signal are shown in Fig. 4.6(b). The observation that they tightly overlap indicates the effectiveness of the transplantation.

4.4 Detecting Anti-Forensics

Our study in Section 4.3 has shown a number of anti-forensic operations that can counteract ENF analysis. In response to these operations, a forensic analyst would devise ways to detect the use of anti-forensic operations, so that a forged



(a) (b)

Figure 4.6: (a) Result of narrowband transplantation around 60Hz; (b) ENF signals extracted from the source signal and from the resulting signal.

audio signal can be identified and rejected as untrustworthy evidence. In this section, we first discuss conditions under which the detection is feasible, and then propose effective detection methods.

4.4.1 Detectability of Anti-Forensic Operations

In order to detect anti-forensic operations, we first provide a mathematical formulation of the anti-forensic operations discussed in Section 4.3. Without loss of generality, the anti-forensic operations proposed therein create a forged audio signal by mixing a bandstopped input signal and a bandpassed alien signal (either real or synthetic). In the frequency domain, the overall anti-forensic operation can be represented as

$$Y(\omega) = e^{-j\alpha\omega}\left[X(\omega)B_s(\omega) + A(\omega)B_p(\omega)\right], \tag{4.2}$$

where $X(\omega)$ is the frequency-domain representation of the original audio signal indexed by the frequency ω (in Hz), $Y(\omega)$ is the resulting audio signal, $A(\omega)$ is the alien signal, $B_s(\omega)$ and $B_p(\omega)$ are the frequency responses of the bandstop filter and the bandpass filter, respectively, and $e^{-j\alpha\omega}$ is a phase shift corresponding to a possible time-domain delay of α. The delay is introduced to avoid boundary conditions due to filtering.

Consider two mutually exclusive cases. For the frequency outside the narrow passband, $|B_s(\omega)| \approx 1$ and $|B_p(\omega)| \approx 0$, and we have

$$\begin{aligned} |Y(\omega)| &\approx |X(\omega)|, \\ \angle Y(\omega) &\approx -\alpha\omega + \angle X(\omega) + \angle B_s(\omega). \end{aligned} \tag{4.3}$$

In practice, both the bandstop and the bandpass filters can be designed as zero-phase or linear-phase. As such, the phase term $\angle B_s(\omega)$ is linear outside the narrowband, and by properly selecting the delay α, the two terms $-\alpha\omega$ and $\angle B_s(\omega)$ can be cancelled out, leading to $Y(\omega) \approx X(\omega)$ outside the narrowband. In other words, the anti-forensic operations basically preserve the host signal outside the narrowband. On the other hand, for the frequency inside the narrowband, we have $|B_s(\omega)| \approx 0$ and $|B_p(\omega)| \approx 1$, and

$$\begin{aligned} |Y(\omega)| &\approx |A(\omega)|, \\ \angle Y(\omega) &\approx -\alpha\omega + \angle A(\omega) + \angle B_p(\omega) \\ &\approx \angle A(\omega) + (\beta - \alpha)\omega \end{aligned} \tag{4.4}$$

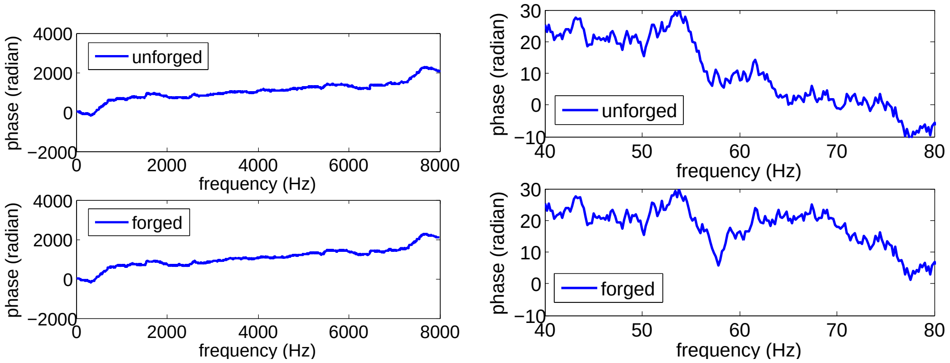
provided that the bandpass filter has linear phase in the narrowband. This suggests that $Y(\omega) \approx e^{(\beta-\alpha)\omega} A(\omega)$, that is, the output signal inside the narrowband resembles the alien signal inside the narrowband with a possible phase shift. If the bandstop and bandpass filters are designed using the same methods, then α and β are similar and thus the phase shift is close to zero. To summarize, overall the proposed anti-forensic operations from Section 4.3 only alter the narrowband and leave no substantial influence outside the narrowband.

To detect anti-forensic operations, a forensic analyst can carry out a likelihood ratio (LR) test to compare the likelihoods of a forged audio signal and an unforged audio signal. Specifically, the analyst evaluates the following likelihood ratio:

$$\begin{aligned} LR &= \frac{P(Y|\text{forged})}{P(Y|\text{unforged})} \\ &= \frac{P(O=o, I=i|\text{forged})}{P(O=o, I=i|\text{unforged})} \qquad (4.5) \\ &= \frac{P(I=i|\text{forged}, O=o)}{P(I=i|\text{unforged}, O=o)}, \qquad (4.6) \end{aligned}$$

where we decompose Y into a pair of (I, O) in (4.5), standing for the inside-narrowband and outside-narrowband components, respectively, and the terms $P(O = o|\text{forged})$ and $P(O = o|\text{unforged})$ are cancelled out in (4.6) since the anti-forensic operations do not affect the host signal outside the narrowband.

For the anti-forensic operations proposed in Section 4.3, the forged narrowband is independent of the signal outside the narrowband. Therefore, the numerator in (4.6) can be written as $P_{I|A}(i)$, standing for the likelihood of observing a narrowband i conditioned that the narrowband is from an alien signal. The denominator, on the other hand, has to account for the dependence of the narrowband on the signal outside the narrowband. Specifically, the denominator can be denoted as $P_{I|X,o}(i)$, which is the likelihood of a narrowband i given that the narrowband is native (i.e., not from another signal) and the signal outside the narrowband is o. In summary,



(a) (b)

Figure 4.7: (a) Comparison of overall phase associated with unforged and forged audio signals; (b) comparison of phase around 60Hz associated with unforged and forged audio signals.

the likelihood ratio is given by $P_{I|A}(i)/P_{I|X,o}(i)$.

From such an analysis, we see that a distinction has to be made between the original audio signal X and the alien signal A in the narrowband, in order to detect anti-forensics operations. This is, however, a challenging task, since the adversary can design the bandstop filter to make the narrowband very "narrow", especially compared to the wide frequency range associated with the much higher sampling frequency. As a result, the characteristics of the original audio signal X and the alien signal A cannot be easily distinguished in the narrowband. To illustrate such a difficulty for the forensic analyst, Fig. 4.7(a) shows the overall phase of an unforged audio signal as well as its forged version, and their difference is hardly noticeable. Zooming into the narrowband as shown in Fig. 4.7(b), we observe that the two versions differ in the narrowband, but it is not straightforward to characterize their statistical difference and to determine which one is forged.

4.4.2 Inter-Frequency Consistency Check

Section 4.4.1 shows that anti-forensic operations can be detected if one can distinguish the two distributions $P_{I|A}(i)$ and $P_{I|X,o}(i)$ in the likelihood ratio. Motivated by this finding, we propose a few ways toward this end.

So far, we have assumed implicitly that a forensic analyst only extracts ENF signals from a given frequency (e.g., the fundamental frequency of 60Hz). In this case, it is reasonable for an adversary to focus on tackling this frequency as well. However, due to the non-linear behavior of electrical circuits, the ENF signal is often present not only at the fundamental frequency, but also at the harmonic frequencies (120Hz, 180Hz, etc) [1]. As such, in order to detect anti-forensic operations, the forensic analyst can perform ENF extraction at more than one frequency, and examine the consistency of multiple ENF estimates. To illustrate this idea, we extract ENF signals from an audio signal at 60Hz and 120Hz, respectively, and the results are shown in Fig. 4.8. Note that these two signals have been normalized with respect to their average values. It can be seen that the two extracted ENF signals highly overlap with each other, and their normalized correlation is 0.97. The power of this check depends on the ENF extraction quality at these harmonic frequencies and is substantially specific to recording conditions. A common observation is that the magnitude of ENF signal at higher harmonic frequencies can be lower, and the host audio signal that interferes with the ENF signal is usually stronger at higher frequencies. As a result, it is usually more difficult to extract reliable ENF signals at higher harmonic frequencies for such consistency check.

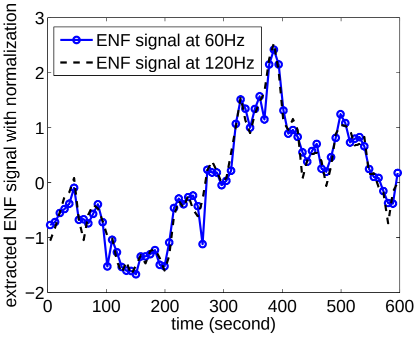
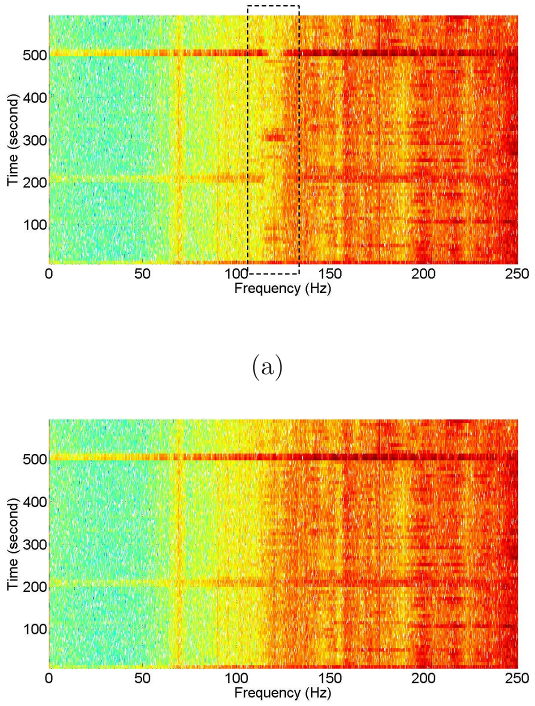


Figure 4.8: Consistency of ENF signals extracted at the fundamental frequency of 60Hz and a harmonic frequency of 120Hz.

4.4.3 Spectrogram Consistency Check

As an adversary performs the anti-forensic operations proposed in Section 4.3, the resulting narrowband often exhibits some kind of inconsistency with the signal outside the narrowband, especially the abrupt boundaries that are easily noticeable around the nominal ENF. Mathematically, this means the value of $P_{I|X,o}(i)$ is small for i that introduces abrupt boundaries, which can be used to indicate the existence of anti-forensics. As an example, consider an adversary that alters the ENF at 120Hz. A typical resulting spectrogram is shown in Fig. 4.9(a), where discontinuity at the narrowband boundaries centered at 120Hz can be clearly noticed. Such inconsistency occurs if the host audio signal and the alien audio signal exhibit salient but unsynchronized temporal variations.

While the spectrogram consistency check is powerful when the signals exhibit inconsistency, automating this check is non-trivial as in reality, a forensic analyst has no *a priori* knowledge about the narrowband range. In order to detect the boundary discontinuity the analyst has to scan the entire frequency range at a fine



(a)

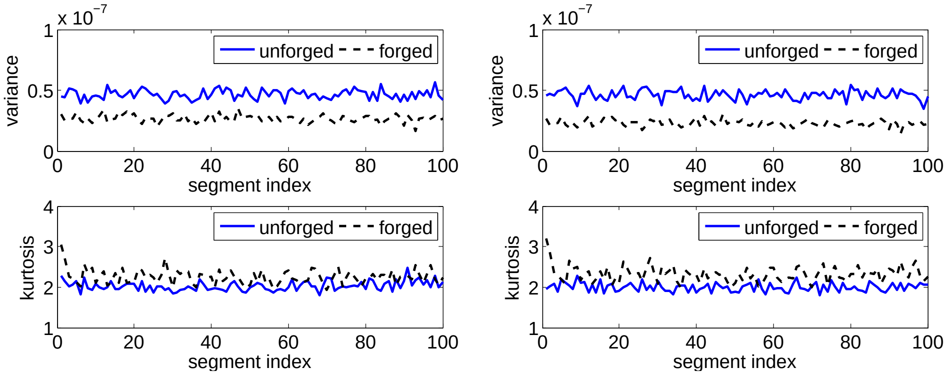
(b)

Figure 4.9: (a) Spectrogram consistency check for a signal with its 120Hz narrowband forged; the obvious inconsistency around 120Hz is highlighted by the dashed box. (b) Spectrogram with an envelope-adjusted narrowband. Notice the inconsistency around 120Hz in Fig. 4.9(a) is no longer visible.

resolution, which demands a high computational complexity.

4.4.4 Reference-based Detection

In Section 4.4.1, we have seen conditions under which anti-forensic operations can be detected. In particular, a forged and an unforged audio signal can be distinguished if their narrowband characteristics are available. Here we consider a special setting called *reference-based anti-forensics detection*, wherein it is assumed that when a query recording's ENF signal is to be authenticated, a reference signal with



(a) Day-1

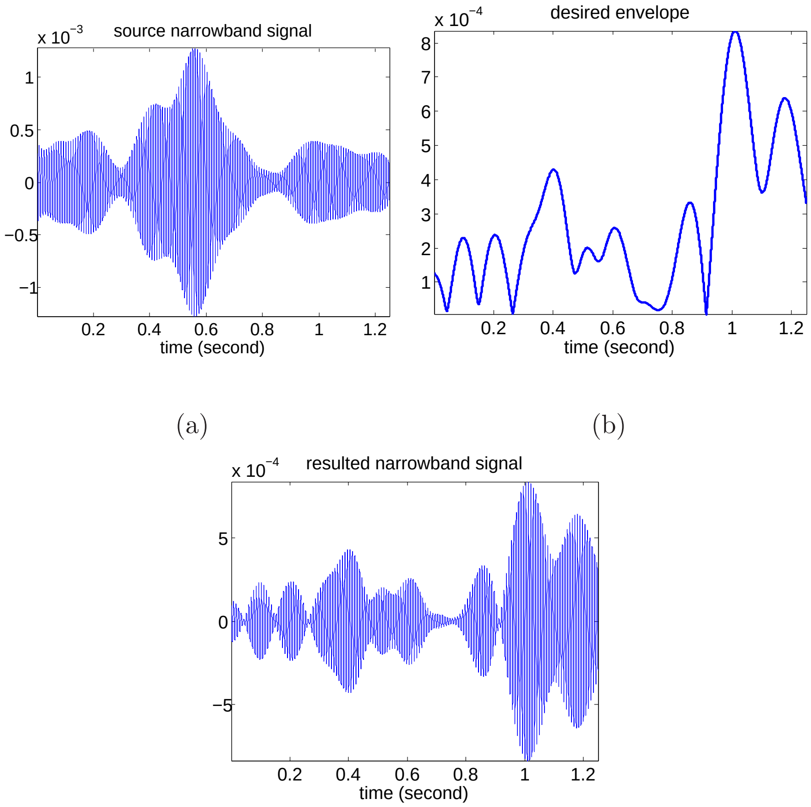
(b) Day-2

Figure 4.10: Variance and kurtosis statistics calculated over 5-second segments on (a) Day 1 and (b) Day 2.

similar ENF sensing conditions is also accessible. Note that this is in contrast to the blind detection method that we have discussed previously. The reference-based setting is feasible in many practical scenarios. For example, if the adversary presents multiple pieces of audio recordings among which some have forged ENF signals, then the remaining unforged audio recordings can serve as the reference signals. As another example, consider an audio file that is used as forensic evidence whose authenticity remains to be determined. A forensic analyst can replicate the recording environment so that the ENF sensing conditions are replicated as well. Note that the reference-based anti-forensics detection can be seen as a resource-augmented detection, and as far as we know, this has not been exploited previously.

In the reference-based anti-forensics detection setting, since the reference signal contains an authentic ENF signal, information about $P_{I|X,o}(i)$ can be learnt from the statistics of the reference signal. Specifically, by writing

$$P_{I|X,o}(i) = P_{I|X}(i)\frac{P(o|i,X)}{P(o|X)},$$



(a)

(b)

(c)

Figure 4.11: (a) The source narrowband signal in time domain; (b) the envelope of the native narrowband signal; (c) the resulting narrowband signal after envelope matching of (a) to (b).

one can detect an anti-forensic operation upon a query audio signal if it leads to a low $P_{I|X}(i)$. To verify this idea, we collect two audio signals recorded on two different days (10 January and 14 January 2012, respectively). The two audio clips were made by playing online streaming via the same speaker and recording using the same microphone. The placement of the microphone and the speaker volume, however, are not strictly controlled on the two days. For a given audio file whose narrowband surrounding 60Hz is denoted by $B(n)$, we divide $B(n)$ into segments of a 5-second duration, and calculate sample statistics for each segment. In particular, we examine the variance that measures how much each sample spreads out from the average value, and the kurtosis that measures the "peakedness" as well as the "tail heaviness" of each sample relative to a normal distribution, defined as

$$\begin{aligned} \mathrm{Var}(B) &= E[(B(n)-\bar{B})^2], && (4.7) \\ \mathrm{Kur}(B) &= \frac{E[(B(n)-\bar{B})^4]}{E^2[(B(n)-\bar{B})^2]}, && (4.8) \end{aligned}$$

respectively, where $\bar{B}$ is the average value of $B(n)$ in a segment. We plot the two statistics corresponding to unforged and forged signals for Day 1 (January 10 2012) and Day 2 (January 14 2012) in Fig. 4.10(a) and Fig. 4.10(b), respectively. We can see that both the unforged and the forged signals have stable statistics on the two days, and unforged and forged signals show noticeably separable statistics values. Therefore, if we are given any of these two unforged recordings as reference, we can detect anti-forensics over the other recording by checking the consistency of the statistics. This idea of reference-based anti-forensics detection can be further augmented by incorporating other useful statistics.

4.5 Concealing Anti-Forensic Traces

Being aware of the anti-forensics detection methods proposed in Section 4.4, the adversary has the incentives to improve the anti-forensic operations. In this section, we explore a few possible methods toward this goal, and discuss their trade-offs.

To cope with the inter-frequency consistency check, the adversary can alter multiple ENF harmonic frequencies. Two issues have to be addressed by the adversary. First, the alteration has to be performed with regard to possible signal quality degradation. This is because altering the ENF signal at higher harmonics involves applying bandstop filtering by the adversary's anti-forensic operations to the audio signal at higher frequencies, which usually has richer content. Second, from a forensic analyst's point of view, as more ENF frequencies are affected, more traces will be left that can be exploited by the reference-based anti-forensics detection. Nevertheless, as discussed in Section 4.4.2, ENF signals generally can only be extracted reliably at lower harmonic frequencies. Around these frequencies, host signal quality degradation is barely noticeable according to our subjective perceptual evaluation. As such, the two issues above are not serious in practice.

4.5.1 Envelope Adjustment

Recall that the anti-forensic operations proposed in Section 4.3 may result in inconsistency on the spectrogram. This is because the forged narrowband may have different temporal magnitude variations. To address this issue, an adversary can try to adjust the envelope of the narrowband, so that the adjusted narrowband has similar temporal variation as the native narrowband. Such adjustment can be done by means of the Hilbert Transform [29]. Specifically, the Hilbert Transform of a real-valued narrowband signal in the form of $b(t) = e(t)\sin(2\pi f_c t + \phi)$ is given by

$$\begin{aligned} H\{b(t)\} &= b(t) + je(t)\sin\left(2\pi f_c t + \phi + \frac{\pi}{2}\right) \\ &= b(t) + je(t)\cos(2\pi f_c t + \phi), \end{aligned} \tag{4.9}$$

which includes a purely imaginary part that is $\pi/2$ phase-shifted from $b(t)$. As a result, the amplitude equals to $|H\{b(t)\}| = e(t)$, where the periodical part $\sin(2\pi f_c t + \phi)$ is no longer present. The envelope adjustment is done by matching the envelopes of the native narrowband and the forged narrowband in the following form:

$$\hat{b}_y(t) = \frac{|H\{b_x(t)\}|}{|H\{b_a(t)\}|} b_a(t), \tag{4.10}$$

where $b_a(t)$ is the source narrowband from the alien signal, and $b_x(t)$ is the narrowband of the original signal. Examples of $b_a(t)$ and $|H\{b_x(t)\}|$ are shown in Fig. 4.11(a) and Fig. 4.11(b), and the resulting narrowband is given in Fig.4.11(c). It is clear that the narrowband from the alien signal has been adjusted with a matched envelope. The spectrogram after envelope adjustment is given by Fig. 4.9(b), which no longer exhibits the spectrogram inconsistency as in Fig. 4.9(a).

Envelope adjustment may cause some loss of fidelity in the forged ENF signal, which can be seen in the following experiment. We perform the narrowband transplantation proposed in Section 4.3.2 on 13 different audio files. Specifically, for each audio file, we extract the narrowband from another arbitrarily chosen file

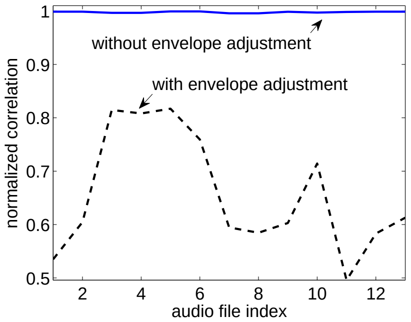


Figure 4.12: Comparison of normalized correlation values with and without envelope adjustment. Note that the normalized correlation has been substantially reduced when envelope adjustment is applied.

and transplant the extracted narrowband into the audio file as described in Section 4.3.2. For these 13 audio files, We first calculate the normalized correlation between the ENF signal present in the alien narrowband and the ENF signal in the forged narrowband. We then perform envelope adjustment and also calculate the normalized correlation between the ENF signal in the alien narrowband and the ENF signal in envelope-adjusted narrowband. As shown in Fig. 4.12, the normalized correlation reduces from a value close to 1 to about 0.6 as a result of the envelope adjustment. That is, the envelope adjustment introduces distortion to the ENF, which suggests that an adversary only has a limited capability of preserving the fidelity of the spectrogram and forged ENF signal at the same time.

4.5.2 Statistics Matching

We have seen in Section 4.4.4 that due to the limited fidelity of ENF forgery, anti-forensic operations may be detectable with the aid of certain statistics from a reference signal. As such, an adversary also has the incentive to match the statistics of a forged signal to those obtained from the reference signal. We have found that the envelope adjustment technique discussed in Section 4.5.1 has the effect of calibrating the variance and kurtosis statistics, as shown in Fig. 4.13, and therefore serves as a technique for counteracting the proposed reference-based statistics matching method as well. However, while the adversary calibrates these two statistics, some other statistics may be affected. For 13 audio recordings, Fig. 4.14 shows the peak magnitude at 60Hz on the FFT result with and without envelope adjustment. We can see that, while the result without envelope adjustment has a wider span, the result with envelope adjustment exhibits a high consistency. This finding can be exploited accordingly by the forensic analyst to detect anti-forensic operations. This phenomenon is fundamental and indicates that some mismatch always takes place if the adversary only has limited knowledge about how ENF is formed in an audio signal. For both forensic analysts and adversaries, it is therefore crucial to acquire a deeper understanding of ENF's underlying mechanism so as to mimic or to scrutinize the fidelity of ENF forgery. The relations between forensic analysts' and adversaries' actions will be discussed in more depth in the next section.

4.6 Understanding the Interplay between Forensic Analyst and Adversary

Summarizing our proposed forensic and anti-forensic operations developed so far, we can see a highly dynamic interaction between the forensic analyst and the

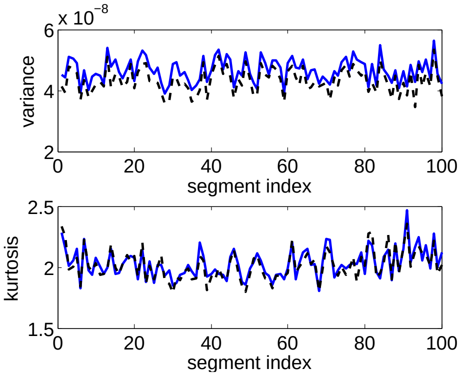


Figure 4.13: Variance and kurtosis statistics matching via envelope adjustment. Solid and dashed curves represent the statistics associated with authentic data and envelope-adjusted data, respectively.

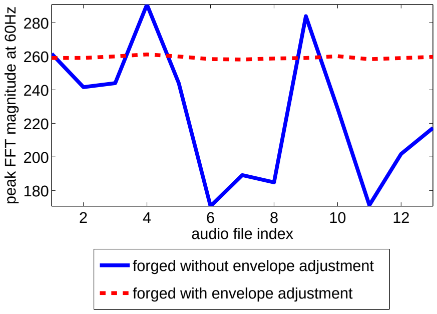


Figure 4.14: Peak FFT magnitude at 60Hz, with and without envelope adjustment. Note that the range of peak FFT magnitude is wider before envelope adjustment and becomes substantially narrower afterwards.

adversary. In this section, we consider such an interaction from two perspectives. The first perspective treats the interaction as an evolutionary process, in which both the forensic analyst and the adversary improve their actions evolutionarily in response to each other's action. We then present a game-theoretic perspective, formulating a game between the forensic analyst and the adversary to highlight their fundamental relation.

4.6.1 An Evolutionary Perspective

In a security context, system attackers and defenders take advantage of vulnerabilities in each other's strategies and advance their own ones. There is always an evolution between the two parties, which has been observed in many practical scenarios such as computer virus v.s. anti-virus competition [50] and the "arms race" for attacking v.s. securing online reputation systems [66]. In a similar spirit, such an evolution can also be observed in ENF analysis, resulting in strategies from simple to complex. As an example, below we list the technical progression from the discussions in earlier sections of this chapter:

1. A forensic analyst extracts ENF at the fundamental frequency (e.g., 60Hz). This is sufficient since the ENF signal is dominant in the narrowband at the fundamental frequency so ENF extraction is accurate, and the forensic analyst does not examine harmonic frequencies that will incur additional complexity.

2. Given the practice in the previous step, an adversary alters the ENF signal at the fundamental frequency using anti-forensic operations proposed in Sec-

tion 4.3 such as removal of the native ENF signals and embedding of a new ENF signal chosen by the adversary.

3. In the presence of the adversary, the forensic analyst is now motivated to extract the ENF signal from other harmonic frequencies to examine the inter-frequency consistency, at the cost of higher complexity.

4. In response to the forensic analyst, the adversary has to make cohesive changes to the ENF signal at higher harmonic frequencies. However, the adversary takes the risk of distorting the host audio signal and has a higher chance of being caught if the forensic analyst applies a reference-based detection.

5. The forensic analyst now has to employ more advanced detection methods at additional costs, such as checking the spectrogram consistency.

6. In response to the forensic analyst's improved detection, the adversary can improve the spectrogram consistency via envelope adjustment. However, this may sacrifice the fidelity of the forged ENF signal.

7. Given that the adversary has addressed the blind detection methods, the forensic analyst can resort to non-blind detection such as checking the signal statistics with reference signals. The means that the forensic analyst can improve his/her capability by resorting to more resources.

8. The adversary now improves the ENF forgery fidelity by matching the statistics at the analyst's disposal. However, we have seen that matching a subset

of the statistics may lead to mismatch of other statistics, and it is difficult to perfectly replicate the authentic ENF formation process.

9. Now the forensic analyst has to seek additional anti-forensics detection methods. The interplay continues.

Evidently, the evolution takes place naturally in a dynamic environment. As this chapter is among the first effort investigating anti-forensics and countermeasures of ENF analysis, we expect that increasingly more sophisticated anti-forensic strategies and countermeasures will emerge and can be characterized by the evolutionary perspective.

4.6.2 A Game-Theoretic Perspective

The interplay between the forensic analyst and adversary in the ENF analysis can be further understood under a game-theoretic framework that is extended from the work by Stamm *et al.* in [63]. Consider the scenario that the forensic analyst extracts the ENF signal at the fundamental frequency (e.g., 60Hz). An adversary present in the system can embed a forged ENF signal as discussed in Section 4.3.1 and Section 4.3.2 upon the audio signal so as to convince the forensic analyst that the audio signal was created at a particular time. As such, for the time information from the extracted ENF signal to be trusted, the authenticity of the ENF signal must first be confirmed by an anti-forensics detector to ensure that no anti-forensic operations have been employed by adversaries.

An anti-forensics detector can be characterized by its structure and performance metrics. In this chapter, we consider a composite construction of anti-forensics detectors. Specifically, consider a total of N individual detectors D_i, $1 \leq i \leq N$, each relying on different signal characteristics to generate a binary output T/F with respect to an input audio signal. Output T (True) means anti-forensics has been performed on the audio signal, and Output F (False) means the opposite (i.e., the audio signal is authentic). An overall anti-forensics detector D_{all} can be constructed using a simple OR-rule:

$$D_{\text{all}} = \begin{cases} T, & \text{if } D_i = T \text{ for any } 1 \leq i \leq N, \\ F, & \text{otherwise.} \end{cases} \tag{4.11}$$

Note that in practice, the detector has constraints on its affordable complexity and the available resources, which determine the individual detectors that can be incorporated into the overall detector. The performance of the detector is measured in terms of its detection probability and false alarm probability. The detection probability is the probability that the detector outputs T given that the anti-forensic operation is performed, and the false alarm probability is the probability that the detector outputs T given that the anti-forensic operation is not performed. There is a common trade-off between these two probabilities of a given detector: the false alarm probability only increases as the detection probability increases. For a total false alarm probability $P_{f,\text{all}}$ allowed for D_{all} that adopts the OR-rule, the forensic analyst's strategy selects and configures individual detectors in terms of their false alarm probabilities, subject to a total false alarm probability equals to $P_{f,\text{all}}$.

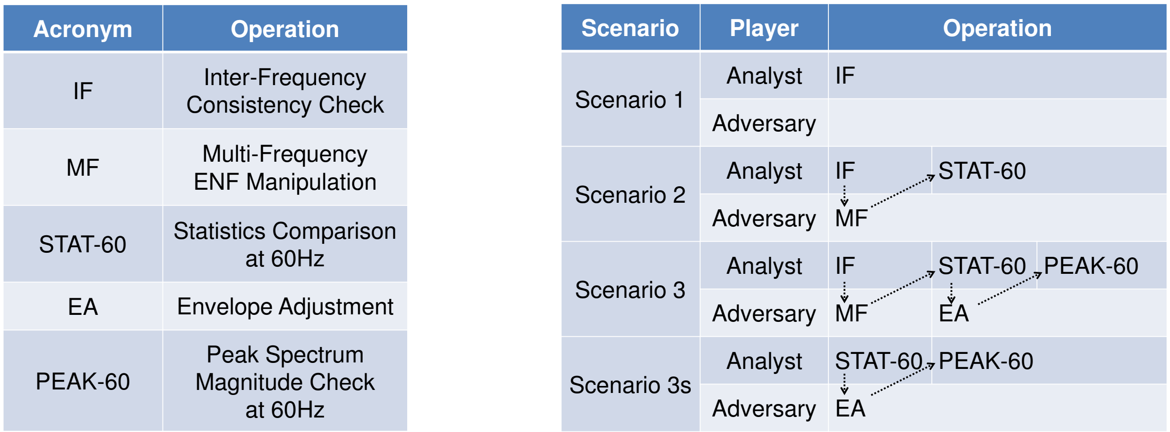
In response to the forensic analyst's anti-forensics detection, the adversary will seek to hide the traces of anti-forensics. Complexity and resource constraints can also be imposed on the adversary's actions, and the adversary has to select his/her strategy under the constraints so that the forensic analyst's detection capability is minimized while the forged ENF signal is maximally preserved. Given a pair of the forensic analyst's and the adversary's strategies, the utility that the forensic analyst will maximize is the total detection probability of anti-forensics $P_{d,\text{all}}$. In contrast, the adversary's utility is to minimize $P_{d,\text{all}}$, with additional penalty when distortion is introduced to the ENF signal that the adversary intends to embed.

The specific operations proposed in Section 4.4 and 4.5 can be studied under the game-theoretic formulation. In terms of the forensic analyst's detector construction, if more strict constraints on complexity and resources are imposed, then the forensic analyst may only use the low-complexity inter-frequency consistency check as the anti-forensics detector. If a higher complexity is permitted, then the spectrogram consistency detector can be incorporated into the overall detector. Furthermore, if the resources accessible to the forensic analyst are enhanced, for example via a reference signal or via an improved understanding of the ENF formation mechanism, then forensic analyst can construct an even more sophisticated detector. On the adversary's side, altering ENF at multiple frequencies is effective against the inter-frequency consistency check, but cannot resist other types of anti-forensics detection. Nonetheless, if higher complexity is allowed for the adversary, he/she can employ envelope adjustment to reduce the anti-forensics detection probability, although at the same time, the forged ENF signal may suffer from distortion. Similar to the forensic analyst, if more resources are available to the adversary, such as an improved knowledge of the ENF formation mechanism, then the adversary can also improve the anti-forensic capability.

4.6.3 Quantitative Evaluation of Representative Scenarios

To establish a concrete and quantitative understanding of the evolutionary and game-theoretic perspectives, we study the scenarios listed in Fig. 4.15 that represent different stages during the ENF "arms race" and can take place in the game-theoretic formulation of ENF forgery. To facilitate the investigation and comparison of players' possible strategies, we first prepare audio recordings to quantitatively test the performance of anti-forensics and countermeasures. Specifically, we collect 100 audio segments by playing online audio streaming via a speaker and recording using a microphone. Each segment is 10-minute long and is 2-minute apart from one another. We consider operations introduced in Section 4.4 and 4.5 that can be performed by the forensic analyst and the adversary, including the inter-frequency consistency check (IF) discussed in Sec. 4.4.2, the statistics comparison around 60Hz (STAT-60) in Sec. 4.4.4, ENF manipulation of multiple harmonic frequencies (MF) in Sec. 4.5, envelope adjustment via Hilbert Transform (EA) in Secs. 4.5.1 and 4.5.2, and the peak spectrum magnitude check around 60Hz (PEAK-60) in Sec. 4.5.2. These acronyms are summarized in Fig. 4.15(a). The dotted arrows in Fig. 4.15(b) represent the causal relations, i.e., one player's action triggers the other player's action.

Scenarios: First, in Scenario 1, the adversary embeds a phony ENF signal at



Acronym	Operation
IF	Inter-Frequency Consistency Check
MF	Multi-Frequency ENF Manipulation
STAT-60	Statistics Comparison at 60Hz
EA	Envelope Adjustment
PEAK-60	Peak Spectrum Magnitude Check at 60Hz

(a)

Scenario	Player	Operation		
Scenario 1	Analyst	IF		
	Adversary			
Scenario 2	Analyst	IF	STAT-60	
	Adversary	MF		
Scenario 3	Analyst	IF	STAT-60	PEAK-60
	Adversary	MF	EA	
Scenario 3s	Analyst	STAT-60	PEAK-60	
	Adversary	EA		

(b)

Figure 4.15: (a) Acronyms of operations and (b) representative scenarios in the ENF forgery game formulation. See Section 4.6.2 for detailed elaborations.

the fundamental frequency of 60Hz, and the forensic analyst performs the inter-frequency consistency check (i.e., the IF detection) in order to detect ENF forgery. The Receiver Operating Characteristic (ROC) curve of the detection, i.e., the relation between the false alarm probability and the detection probability, is shown in Fig. 4.16. The nearly perfect detection performance suggests that the inter-frequency ENF discrepancy can effectively detect ENF manipulations at a single frequency.

Scenario 2 considers the further interaction when the adversary performs the ENF manipulation of multiple harmonic frequencies in order to counteract the forensic analyst's inter-frequency consistency check. Assume that the forensic analyst has access to a reference signal with similar statistics, then the forensic analyst can perform the STAT-60 detection to verify the ENF signal's statistics present at 60Hz. Fig. 4.17(a) shows the substantial performance drop of the IF detection due to the

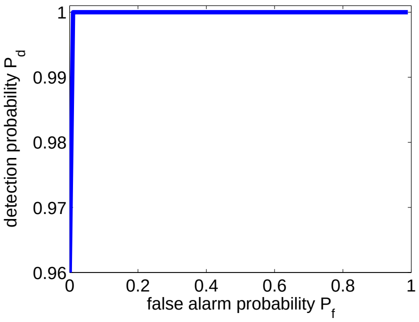
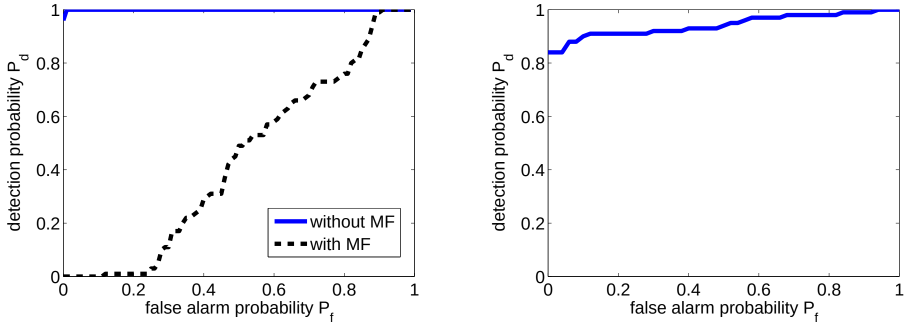


Figure 4.16: ROC curve of IF detection that performs inter-frequency consistency check.

multi-frequency ENF manipulation, and it is clear that the inter-frequency consistency check is no longer effective in this scenario. However, STAT-60 the compares the statistics at 60Hz remains discriminative as shown in Fig. 4.17(b), and therefore if a composite detector is constructed using IF and STAT-60, STAT-60 should play a dominant role and the forensic analyst should always assign the available false alarm probability to STAT-60.

In Scenario 3, the adversary further counteracts STAT-60 by applying envelope adjustment via Hilbert Transform. As discussed in 4.5.2, envelope adjustment can match the statistics used by STAT-60, which is also confirmed in Fig. 4.18(a), where one can see that the STAT-60 detection essentially becomes a random guess in the presence of envelope adjustment. On the other hand, however, the downside to envelope adjustment from the adversary's perspective is that the forged ENF signal may be distorted as shown in Fig. 4.12. A feasible compromise available to the adversary is to control the strength of envelope adjustment by, for example linear



(a)

(b)

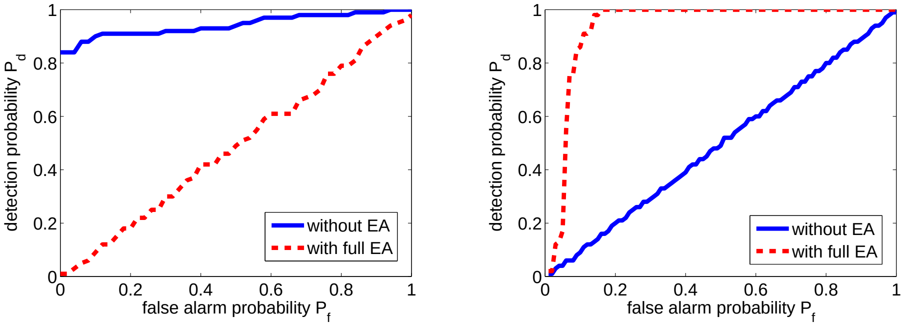
Figure 4.17: (a) ROC curve of IF detection, with and without the multi-frequency ENF manipulation operation (MF); (b) ROC curve of STAT-60 detection.

mixing, i.e.,

$$\hat{b}_{y,\alpha}(t) = \frac{\alpha|H\{b_x(t)\}| + (1-\alpha)|H\{b_a(t)\}|}{|H\{b_a(t)\}|} b_a(t),$$

where $0 \leq \alpha \leq 1$ denotes the strength of envelope adjustment. It can be seen that a higher α makes the adjusted envelope more similar to that of the native narrowband. A higher α also introduces more distortion to the forged ENF signal as discussed earlier. Now, in response to the practice of envelope adjustment, the forensic analyst applies the PEAK-60 detection that scrutinizes the peak spectrum magnitude at 60Hz, whose ROC curves with and without full envelope adjustment ($\alpha = 1$) are shown in Fig. 4.18(b). We can see that PEAK-60 behaves as a random guess in the absence of envelope adjustment, but becomes discriminative in the presence of envelope adjustment.

Nash Equilibria and Optimal Strategies: We now consider the optimal strategies of the forensic analyst and the adversary as well as the resulting forensic and anti-forensic performance in Scenario 3. Here, the notion of strategy optimality



(a) (b)

Figure 4.18: ROC curve of (a) STAT-60; (b) PEAK-60, with and without envelope adjustment (EA).

refers to a Nash Equilibrium, namely, the status in which no player can increase his/her own utility via unilateral strategy changes. As Scenario 3 involves three detectors (IF, STAT-60, and PEAK-60), the forensic analyst's strategy is to configure the composite detector by setting the false alarm probabilities of individual detectors subject to the total false alarm probability. This strategy has two degrees of freedom and is more difficult to observe directly. To gain some useful insights, we first consider a simplified version Scenario 3s, which does not involve multiple harmonic frequencies (i.e., inter-frequency consistency check and multi-frequency ENF manipulation are not used).

In Scenario 3s, for an assigned value of $P_{f,\text{all}}$, the forensic analyst searches for possible values of $P_{f,\text{PEAK}-60}$ that can be combined with a corresponding $P_{f,\text{STAT}-60}$ to yield a total false alarm probability of $P_{f,\text{all}}$. On the other side, the adversary considers different values of envelope adjustment strength α, subject to any fidelity constraint on the forged ENF signal. that can be mapped into a corresponding constraint on α. Since the goal of the forensic analyst is to detect anti-forensics and the goal of the adversary is to evade the detection, we choose the utility function of the forensic analyst as the overall detection probability $P_{d,\text{all}}$, and the utility function of the adversary as $-P_{d,\text{all}}$. It can be seen that this is a zero-sum game setting, for which the Nash Equilibrium (NE) of this game is given by the min-max (or equivalently max-min) solution. That is,

$$\begin{aligned}(P^{*}_{f,\text{PEAK}-60}\ ,\ \alpha^{*}) = &\ \arg \max_{P_{f,\text{PEAK}-60}} \min_{\alpha} P_{d,\text{all}} \\ = &\ \arg\min_{\alpha} \max_{P_{f,\text{PEAK}-60}} P_{d,\text{all}}, \\ &\ \text{subject to } \alpha \leq \alpha_T\end{aligned} \tag{4.12}$$

for some α_T that upper-bounds the envelope adjustment strength and therefore controls the fidelity of the forged ENF signal.

Fig. 4.19 illustrates the utility function $P_{d,\text{all}}$ for $P_{f,\text{all}} = 10\%$ with respect to different values of $P_{f,\text{PEAK}-60}$ and α. We have several observations: 1) $P_{d,\text{all}}$ generally decreases as α increases, but certain "rebounds" can also be seen for larger values of $P_{f,\text{PEAK}-60}$. The decreasing trend of $P_{d,\text{all}}$ can be attributed to the fact that α's increase reduces the STAT-60 detector's discriminative capability, which may not be well compensated by the improved detection capability of PEAK-60 until a minimum value of $P_{d,\text{all}}$. After that, PEAK-60 begins to compensates for the lost detection capability of STAT-60 and thus $P_{d,\text{all}}$ increases. 2) $P_{d,\text{all}}$ generally increases as $P_{f,\text{PEAK}-60}$ increases. For larger α, this is because PEAK-60 is more discriminative than STAT-60, and even for smaller α when PEAK-60 behaves nearly as random guess, the detection probability of STAT-60 may not increase as rapidly with its false probability as of PEAK-60, and therefore incorporating PEAK-60 by choosing a higher $P_{f,\mathrm{PEAK-60}}$ still increases the overall detection probability.

For the illustrative example, note that when there is no constraint imposed on the envelope adjustment strength α, a particular Nash Equilibrium (NE) can be found as $(P^*_{f,\mathrm{PEAK-60}}, \alpha^*) = (10\%, 80\%)$. That is, the equilibrium takes place when the forensic analyst assigns all of its false alarm probability to PEAK-60, and the adversary uses a large but not maximal strength when adjusting the envelope. In case the envelope adjustment strength is upper-bounded by $\alpha_T < 80\%$, the Nash Equilibrium becomes $(10\%, \alpha_T)$.

For the unconstrained case, we show the NE ROC curve in Fig. 4.20(a), which can be obtained by varying the value of $P_{f,\mathrm{all}}$ and finding the corresponding Nash Equilibrium and $P_{d,\mathrm{all}}$. It can be seen that the detection performance is lower than the solid curve in Fig. 4.18(a) that represents the optimal performance of STAT-60 when no adversarial operation is involved. It is also lower than the dashed curve in Fig. 4.18(b), which is the optimal performance of PEAK-60 when the full application of envelope adjustment is known in advance. Such degradation in identification performance comes from the manipulation of the adversary; nevertheless, the detection performance is retained to a large extent if the forensic analyst adheres to the Nash Equilibrium. Another observation is shown in Fig. 4.20(b) that the Nash Equilibrium strategy for the adversary, i.e., the envelope adjustment strength α, decreases as $P_{f,\mathrm{all}}$ increases. This can be understood from Fig. 4.18, where one can see that STAT-60 without envelope adjustment exhibits a higher detection performance than PEAK-60 with envelope adjustment in the low false alarm probability

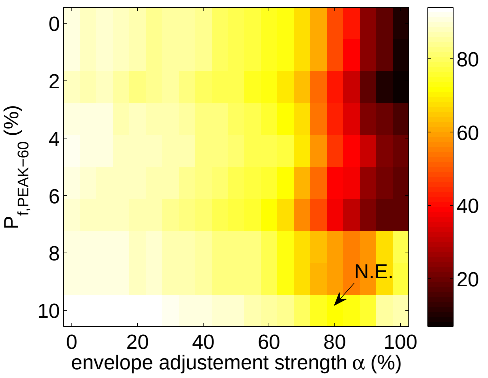
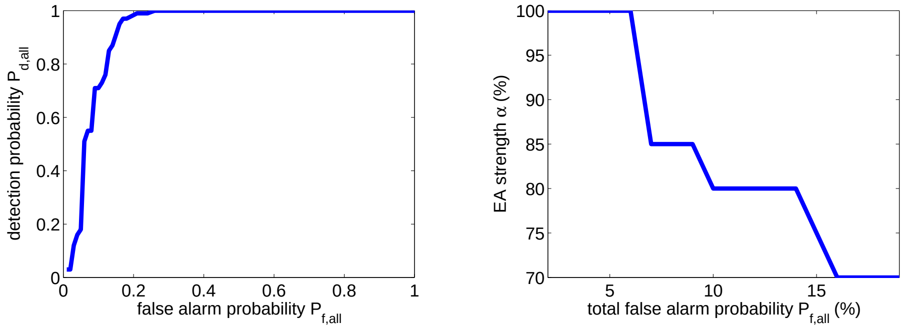


Figure 4.19: Overall detection probability $P_{d,\mathrm{all}}$ as utility function for $P_{f,\mathrm{all}} = 10\%$, evaluated with respect to joint selection of $P_{f,\mathrm{PEAK-60}}$ and α. An unconstrained NE can be found at $(P_{f,\mathrm{PEAK-60}}, \alpha) = (10\%, 80\%)$.

regime. Therefore, when $P_{f,\mathrm{all}}$ is small, the adversary's better strategy is to motivate the forensic analyst to use PEAK-60 by maximizing the envelope adjustment strength. As $P_{f,\mathrm{all}}$ increases, the detection performance of PEAK-60 improves, the above strategy becomes less effective, and the adversary naturally reduces the envelope adjustment strength.

Scenario 3 essentially shares the Scenario 3s' properties. Since its utility function involves three dimensions and is more difficult to visualize, we just plot its NE ROC curve in Fig. 4.21, shown jointly with the NE ROC curve of Scenario 3s for the sake of comparison. We can see that the two ROC curves essentially overlap, which implies that the inter-frequency consistency check and the multi-frequency ENF manipulation do not play meaningful roles at the Nash Equilibrium. Other observations, especially that the envelope adjustment strength decreases as the total false alarm probability increases, are also valid in Scenario 3.



(a) (b)

Figure 4.20: (a) NE ROC curve of Scenario 3s; (b) The optimal envelope adjustment strength α^* at NE with respect to total false alarm probability $P_{f,\text{all}}$.

In summary, our quantitative evaluations of representative scenarios presented in this section provides an important understanding on the optimal strategies of the forensic analyst and the adversary. In particular, we can see that the adversary can effectively reduce the detection performance by properly selecting the envelope adjustment strength, but in the meantime, the forensic analyst's optimal configuration of the composite detector can minimize such a performance degradation. Also note that the game-theoretic analysis here is generic in nature, and it can be extended to other scenarios as well when new anti-forensic operations and countermeasures become available.

4.7 Chapter Summary

The time stamp based on the electrical network frequency (ENF) has been shown to be a promising tool for digital recording authentication. In this chapter, we examined the robustness of this time stamp against anti-forensics under

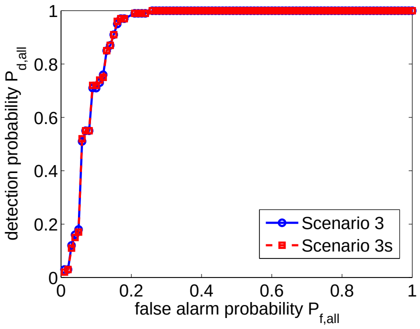


Figure 4.21: NE ROC curves of Scenario 3 and Scenario 3s, which are essentially overlapped.

adversarial environments. We have investigated anti-forensic operations that can remove and alter the ENF signal present in a host audio signal. We have developed a mathematical framework for ENF modification, which not only entails the effectiveness of ENF modification and challenges of anti-forensics detection, but also motivates detection methods from a forensic analyst's point of view. Concealment techniques in response to the anti-forensics detection are further proposed and their corresponding trade-offs are discussed. To understand the dynamic nature of the forensic analyst-adversary interplay, we have developed an evolutionary perspective and a game-theoretic perspective, which can be used to characterize a wide range of actions that may take place. Representative scenarios that involve different actions have also been quantitatively evaluated and the optimal strategies have also been derived.

As this chapter has established a methodology for studying the robustness of ENF-based time stamps, our future work will include more experiments that cover a variety of testing conditions, geographic areas and recording devices. Equally important is to develop a deeper understanding of the ENF formation mechanism as well as individual anti-forensic operations and countermeasures. Certain physical means, such as electromagnetic shielding or the limited frequency response of microphones, may also affect the presence of ENF signals and warrant more research. In light of the potential employment of ENF analysis for digital recording authentication, we envision that its robustness will receive increasing attention, and research along this direction will contribute to more reliable time stamp schemes based on ENF analysis.

CHAPTER 5

Camera Unit Identification using Low-bit-rate Video

5.1 Chapter Introduction

Pocket-sized digital cameras and cell-phones with cameras have become popular and generated a large amount of digital images and videos. Compared to images, videos can capture more visual information, and therefore is an ideal format for recording rich and dynamic content.

Accompanying the growing importance of digital videos, concerns regarding their origin and authenticity have been raised and are receiving increasing attention. A systematic study of *digital video forensics* that answers different questions about a video's acquisition and processing history is important in order to establish the trustworthiness of digital videos. Several previous works on video forensics considered the identification of source devices and tampering operations. In [12], Chen *et al.* extended the source camera identification technique based on the Photo-Response Non-Uniformity (PRNU) [24] from image to video. McCloskey [46] proposed to take into account the influence of video content on the achievable performance of [12]. On tampering detection, Wang and Farid [74] demonstrated that frame insertion or deletion that are usually involved in video forgery form forensic traces and therefore can be detected. Luo *et al.* [45] showed that MPEG compression introduces different block artifacts into different types of frames, which can be used to detect video recompression.

In this chapter, we examine the source camera identification problem, with a focus on *cell-phone cameras*. We focus on cell-phone cameras because more cell-phones are now equipped with the video recording capability, and we foresee that more videos will be generated by cell-phones in the future owing to their superior convenience. Previous works such as [39, 44] have developed and enhanced the methodology of source camera identification by means of the PRNU [24] which we will review shortly. These works considered the case when *still images* from the camera under investigation are used for PRNU estimation and matching. This methodology is extended in [12] to use *videos*, and the reported accuracy is promising when the test video is long enough. However, as also noticed in [46], the task of source camera identification using videos is more challenging than the image counterpart due to the degraded visual quality of videos. This problem is even more serious when we consider videos generated by cell-phone cameras that suffer from much stronger compression. Nevertheless, the rich temporal information in videos can help, if properly exploited, to achieve more accurate source camera identification.

As a video is composed of multiple frames, how each frame should be used to jointly estimate the PRNU deserves careful exploration. In this chapter, we study the effect of video compression, and show that the reliability of frames for PRNU estimation can be considerably different, attributed to different levels of compression. We propose new mechanisms for PRNU estimation that leverage such a difference, and show that more accurate source camera identification can be achieved with fewer frames used.

5.2 PRNU for Source Camera Identification

We review the basic principles of source camera identification based on PRNU. For a more detailed discussion, please refer to [24]. The manufacturing imperfections of charge-coupled device (CCD) and complementary metal-oxide semiconductor (CMOS) sensors result in slight variations of the sensitivity of sensors to the incident light. The pattern of sensitivity variation, commonly referred to as the Photo Response Non-Uniformity (PRNU) [24], can be seen as the "fingerprint" unique to individual imaging devices. It has been shown in [24] that, by applying a denoising filter on the image $\mathbf{F}$, the difference between $\mathbf{F}$ and its denoised version can be approximated by $\mathbf{V} = \mathbf{F}\mathbf{K} + \mathbf{M}$, where $\mathbf{V}$ is referred to as the noise residual, $\mathbf{K}$ is the PRNU pattern matrix that captures the variation pattern of sensor sensitivity, and $\mathbf{M}$ is the modeling noise that accommodates various noise sources, including shot noise, dark current, read-out noise, quantization and compression noise, and the imperfection of the denoising filter. Please be informed that all multiplication operations throughout this chapter are element-wise.

For source camera identification using output images, it is usually assumed that N images taken by the camera under investigation are available for PRNU estimation. When the modeling noise $\mathbf{M}$ is assumed as white Gaussian with per-pixel variance identical across all the images, a maximum-likelihood estimate of $\mathbf{K}$ can be derived as:

$$\hat{\mathbf{K}} = \frac{\sum_{i=1}^{N} \mathbf{V}_i \mathbf{F}_i}{\sum_{i=1}^{N} (\mathbf{F}_i)^2}, \tag{5.1}$$

where $\mathbf{V}_i$ and $\mathbf{F}_i$ are the ith noise residual and ith image, respectively [24].

The typical setting of source camera identification assumes the camera under investigation is available. To match test images against this camera, a training procedure is performed first to obtain a *reference PRNU*. Ideal training images are those with smooth content and high yet unsaturated luminance. Then a PRNU estimate from the test image is calculated using Eq. (5.1) and compared against the reference PRNU. A popular sub-optimal similar metric between two PRNU matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ is the Normalized Cross-Correlation (NCC) given by

$$\mathrm{NCC}(\mathbf{S}_1, \mathbf{S}_2) = \frac{(\mathbf{S}_1 - \bar{\mathbf{S}}_1) \otimes (\mathbf{S}_2 - \bar{\mathbf{S}}_2)}{\|\mathbf{S}_1 - \bar{\mathbf{S}}_1\| \|\mathbf{S}_2 - \bar{\mathbf{S}}_2\|},$$

where $\otimes$ denotes the dot product, and $\bar{\mathbf{S}}_1$ and $\bar{\mathbf{S}}_2$ are the average value of $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively. A correlation matrix $\mathbf{C}$ can be obtained where $C(i, j)$ is the NCC value between $\mathbf{S}_1$ and $\mathbf{S}_2$ when $\mathbf{S}_2$ is shifted by (i, j). Another PRNU similarity metric that compensates for the camera-specific NCC range is called the Peak to

Correlation Energy (PCE), defined as

$$\mathrm{PCE}(\mathbf{S}_1, \mathbf{S}_2) = \frac{(n - |\mathcal{N}_{\mathrm{peak}}|)C_{\max}^2}{\sum_{(i,j)\notin\mathcal{N}_{\mathrm{peak}}} C(i,j)^2},$$

where $C_{\max} = \max_{i,j} C(i,j)$, $\mathcal{N}_{\mathrm{peak}}$ is a small neighborhood surrounding the shift corresponding to $C_{\max}$, and n is the size of $\mathbf{S}_2$. PCE characterizes if the maximum correlation is much higher than the average correlation, or in other words, if there is a peak in the correlation matrix. We adopt the PCE metric in this chapter.

PRNU-based source device identification using output videos has been studied in previous works [12] and [46]. Particularly, in [12], PRNU is utilized to determine if two video clips come from the same source camcorder. The main idea is to treat each frame as one image in a video consisting of N frames, and then apply Eq. (5.1) to obtain an estimate based on the multiple frames, *i.e.*, the entire video. It is advised in [12] that each frame be treated equally mainly to reduce the complexity of implementation. The authors reported that source camcorder can be identified as long as the video is sufficiently long. In [46], the method described above is examined with special attention to the influence of video content. It was observed that edges can be mistaken as noise by the denoising filter, which is further amplified if frames in the video are highly correlated. It is proposed in [46] to assign higher weights to pixels in smooth areas to alleviate this problem, which actually shares a similar spirit with other image-based PRNU estimation techniques such as [39].

5.3 Compression Effect on PRNU Estimation

Most of cell-phone cameras today support low-bit-rate video coding standards MPEG-4 AVC/H. 264. The typical resolution ranges from 320 × 240 to 480 × 352 pixels, and the bit rate may vary between 300 to 1000 kbps. Such strongly-compressed videos are generated in order to meet a more stringent storage-space constraint and to reduce the transmission effort. Strong compression may lower the accuracy of PRNU estimation, as it creates blocking artifacts and coarsely quantized intensity levels, and eliminates a significant amount of content detail that carry the PRNU-induced noise.

We take an empirical approach to understand the impact of compression on PRNU estimation, in particular, if different frames have different reliability for PRNU estimation [15]. As it is a non-trivial task to calculate the frame quality without the uncompressed video for reference, we judge the frame reliability in terms of their correlation with the reference PRNUs. We collect 5 recently-released cell-phones with video recording capability as listed in Table 5.1. Twenty videos that contain indoor and outdoor scenes of 30 seconds are taken with each camera. Interestingly, we find that all frames are either I- or P-frames, and no B-frame is found. We obtain the reference PRNUs of all these cameras according to the procedure recommended in Sec. 5.4. The sequence of frame type of each video can be represented as $\{I, P_1, P_2, P_3, P_4, \ldots, I, P_1, P_2, P_3, P_4, \ldots, I, \ldots\}$. The PRNU of each test video can be estimated with the subset of frames corresponding to the same symbol (*i.e.*, the same offset from I-frames).

Table 5.1: Cell-phone cameras used in our experiment

Index	Model	Format	Resolution
1	RIM Blackberry 9530	3GP	480 × 352
2	Sony Ericsson W705a	MP4	320 × 240
3	Motorola Cliq	3GP	352 × 288
4,5	Apple iPhone 4 (×2)	MOV	568 × 320

For each camera, the PCE value averaging over 20 videos with matched against reference PRNU is shown in Fig. 5.1. The PCE value is much higher (about twice) when the PRNU is estimated using I-frames, but the difference in PCE between different subsets of P-frames is not obvious. That is, the PRNU extracted from I-frames are more correlated to the reference PRNU than those from P-frames, which implies that I-frames are more reliable than P-frames for PRNU estimation. In the meantime, the average PCE values associated with P_1, P_2, P_3, and P_4 have similar values of $31.8, 30.6, 32.7, 32.0$, respectively, indicating that P-frames with different offsets have similar reliability for PRNU estimation.

5.4 Reference PRNU Estimation

In order to perform resilient matching between the reference PRNU and the PRNU from test videos, it is crucial to obtain reliable reference PRNUs in the training process. As compression poses a critical impact on PRNU estimation as shown in Sec. 5.3, it is reasonable to favor I-frames if enough I-frames are available. Besides, since the compression under our consideration is strong, various noise sources may

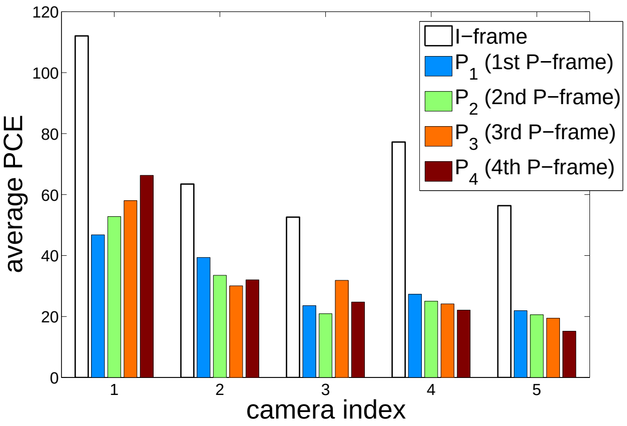


Figure 5.1: Average PCE for different offsets from I-frames.

be dominated by compression noise that is highly content-dependent. One should avoid the use of videos with (nearly) static content otherwise the overall modeling noise associated with different frames in a video will be unfavorably correlated and cannot be easily removed through frame averaging.

These observations motivate us to use multiple short videos, instead of one long video, to obtain the reference PRNU. Specifically, a total of N short videos (shorter than 1 second) that contain smooth and bright scenes are first collected, and then the first frame of each video will be used to jointly estimate the reference PRNU. Since practically the first frame in each video is an I-frame, there are as many I-frames as the number of training videos available for reference PRNU estimation. Moreover, because these I-frames are from different videos, it is expectable that they will have lower correlation with one another.

We compare this mechanism of reference PRNU estimation with two alternatives: 1) using the first P-frames (*i.e.*, the second frame in a video) from multiple videos and 2) using a long video with static content. We refer to these three mech-

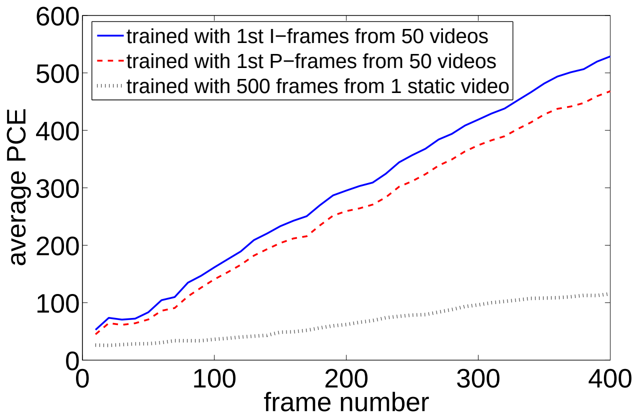


Figure 5.2: Comparison of different mechanisms for reference PRNU estimation, in terms of the achievable PCE value for different test-video frame numbers. Blackberry 9530 is used.

anisms as $\mathcal{M}_I$, $\mathcal{M}_P$, and $\mathcal{M}_L$, respectively. For $\mathcal{M}_I$ and $\mathcal{M}_P$, 50 short videos are used to estimate the reference PRNU. For $\mathcal{M}_L$, a long video with 500 static frames is used to estimate the reference PRNU. In Fig. 5.2, we show for the three mechanisms the PCE values averaging over 20 test videos with respect to different frame numbers from the test video. One can see that $\mathcal{M}_I$ is consistently superior to $\mathcal{M}_P$, which increases as more frames from the test video are used. On the other hand, estimating the reference PRNU using a long but static video is much less effective. If the reference PRNU is obtained in such a way, then even if much more frames in the test video are used, then correlation between the test-video PRNU and the reference PRNU is still much smaller.

5.5 Efficient PRNU Matching by Frame Reordering and Weighting

We have shown that I-frames extracted from videos are more reliable than P-frames for PRNU estimation. Nevertheless, the average PCE value when all frames are used is 300.9, much higher than the average PCE value of 72.3 if only I-frames are used. It is therefore reasonable to use all the frames in a video to obtain a PRNU estimate, and this is in line with the conclusion made in [12]. Two issues, however, need to be addressed more carefully. First, using all the frames in a video can be prohibitively time-consuming, since all frames have to go through a denoising process with non-negligible complexity to extract the frame-wise PRNU. Besides, since I-frames and P-frames have distinct reliability, they should be treated differently when combined for PRNU estimation.

To address the first issue, if the number of frames that can be processed in PRNU estimation is limited, a reasonable choice is to first use more reliable frames, *i.e.*, I-frames. This is feasible in terms of video decoding complexity since I-frames are at the beginning of the Group of Picture (GOP) and can be easily located. In this chapter, we assume that information required to decode the subsequent P-frames are stored after an I-frame is completely decoded, so that the decoding of P-frames can be performed without re-decoding the I-frames. For the second issue, by allowing the ith frame has its modeling noise variance of σ_i^2, we can generalize Eq. (5.1) as

$$\frac{(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} \mathbf{V}_i \mathbf{F}_i)}{(\sum_{i=1}^{N} \frac{1}{\sigma_i^2} (\mathbf{F}_i)^2)},$$

which indicates that a frame should be assigned a weight inversely proportional to its modeling noise variance. We assume that all I-frames have the same modeling noise variance of σ_I^2, and all P-frames have the same modeling noise variance of σ_P^2. Since videos generated by cell-phones are strongly compressed, σ_I^2 and σ_P^2 are mainly determined by the level of compression noise, and therefore should be directly related to the signal-to-noise ratio (SNR) of each frame type. Estimating the SNR using only the compressed video is in general a difficult task [7]; in this chapter, we arbitrarily take $\sigma_P^2 = 2\sigma_I^2$, or equivalently assign weights 2 and 1 to I-frames and P-frames, respectively. Please be reminded that this setting is merely to demonstrate that proper weighting may improve PRNU estimation.

We compare the sequential frame parsing (*i.e.*, reading frames from the beginning of the video in a sequential manner), the proposed frame reordering mechanism with equal weights, and the proposed frame reordering mechanism with the $2:1$ weights. Fig. 5.3 shows the PCE values for these three mechanisms, averaging over totally 100 videos from 5 cameras. One can see that 1) with more frames, the difference between the match and mismatch cases becomes more obvious; 2) the frame reordering mechanism significantly increases the PCE values, especially when the frame number is smaller; 3) for all the frame numbers, the $2:1$ weights assigned to I-frames and P-frames create additional increase in PCE. Note that these two mechanisms do not increase the PCE in the mismatch case.

We also compare these mechanisms in terms of their source camera identification accuracy. The Receiver Operating Characteristic (ROC) curves for the three mechanisms for two frame numbers 100 and 300 are shown in Fig. 5.4 and 5.5, where the horizontal axis is the false alarm rate and the vertical axis is the detection rate.

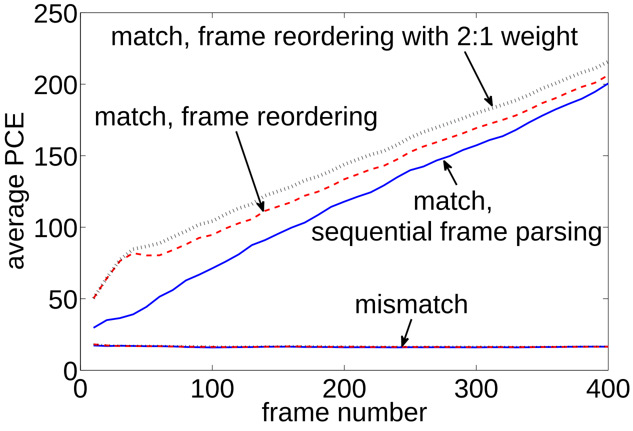


Figure 5.3: Average PCE value with respect to different number of frame.

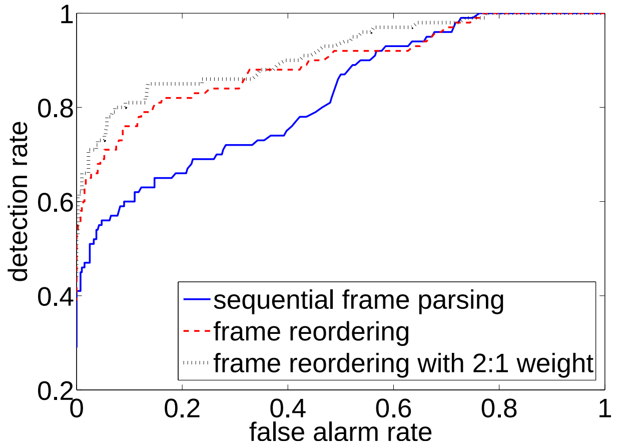


Figure 5.4: ROC curve with 100 frames for PRNU estimation.

One can see that with an increased number of frames, the accuracy is improved for all the three mechanisms. Frame reordering increases the accuracy especially for a smaller number of frames, and further improvement can be obtained by assigning higher weights to more reliable frames. It is also noteworthy that frame ordering and unequal weighting have a complimentary nature: the former is advantageous if only a limited number of frames can be processed, while the latter is more useful if more frames are available.

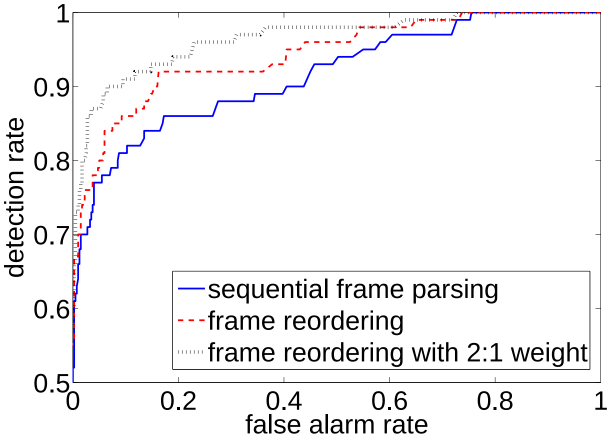


Figure 5.5: ROC curve with 300 frames for PRNU estimation.

5.6 Chapter Summary

In this chapter, we explore the impact of compression on source camera identification using the Photo-Response Non-Uniformity (PRNU) extracted from compressed videos. We consider videos generated by cell-phone cameras, which are strongly compressed to reduce the storage and transmission requirement. Although the authors in [12] stated that each frame in a video should be treated equally, we find that different frame types (I and P) actually have different levels of reliability for PRNU estimation. Motivated by this observation, we propose an effective mechanism for estimating the reference PRNU pattern. Moreover, we show that by reordering and weighting the frames in a video according to their reliability, we can achieve more accurate source camera identification with fewer frames used.

CHAPTER 6

Empirical Frequency Response for Digital Image Forensics

6.1 Chapter Introduction

In the past decade, due to the widespread popularity of digital cameras and online image hosting services, a large number of images have been generated and distributed. At the same time, the advent of various image editing software packages has made altering the image content easier even for novice users. Since the authenticity of digital images impacts on how we use it, content integrity has become an important forensic issue. For a given image, one may ask if it has been tampered or manipulated and further by what *type* of tampering operation. This chapter focuses on the latter question and presents a framework to determine the *type* of tampering operation that has been performed.

Prior work fall into two main categories. In the first category, methods have been proposed to detect resampling [58], JPEG compression [43], and Gamma correction [20], by extracting certain salient features that would help distinguish such tampering from unprocessed images. Although these methods can be employed to identify the type and the parameters of the tampering operation, an exhaustive search over a pool of operations is required to detect tampering and to identify the type of tampering operation. Therefore, there is a strong need for universal technique to detect and identify tampering.

In the second category, classifier-based approaches to detect image tampering were proposed in [4] [23], where features based on analysis of variance [4] and higher order wavelet statistics [23] have been used. In [69], a framework was proposed by modeling tampering as a combination of a linear and shift-invariant (LSI) and a non-LSI part. The authors present methods to estimate the LSI part of manipulation operation and compare the estimate to an identity transform to detect tampering. These work aim to just *detect* tampering and therefore focus on answering whether the given image was tampered or not, and are not for identifying the *type* of tampering.

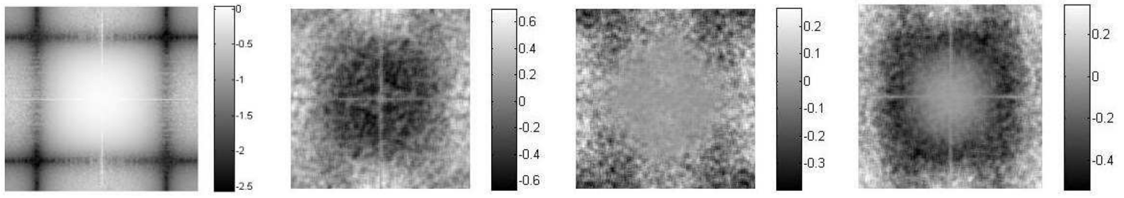
In this work, we propose a framework based on the Empirical Frequency Response (EFR) that aims to identify the manipulation type. We show that many classes of LSI or non-LSI image processing operations, such as resampling, JPEG compression, and non-linear filtering, exhibit distinctive patterns in their EFRs. Theoretical reasoning supported by experimental results also verifies the effectiveness of this method for identifying the type of a tampering operation.

We also find that the EFR potentially can be used for other applications. Specifically, the EFR has dependency on the camera model used to generate the image, and such dependency can thus be leveraged to identify the camera model. Our study also shows that the dependency is a function of the frequency region, which suggests the need for a proper selection of the frequency region.

This chapter is organized as follows. We define the Empirical Frequency Response (EFR) in Section 6.2 and show distinctive EFRs. The results on using the EFR as a tampering analysis tool are discussed in Section 6.3. The application of EFRs for camera model identification is presented in Section 6.4. Since the EFR is, in fact, not readily available in practice, we discuss methods to estimate EFR in Section 6.5 just based on the output image, and propose approches to improve the accuracy. We summarize this chapter in Section 6.6.

6.2 The Empirical Frequency Response

It is well known that linear and shift-invariant (LSI) systems can be characterized by their frequency responses. For example, a 3 × 3 average filter has a 2-D sinc-like frequency response as shown in Fig. 6.1(a) and the frequency response of an identity system whose output equals to the input is flat. However, image processing operations are often non-LSI and input-independent frequency response is not defined for such systems. In this chapter, we represent such manipulations using the Empirical Frequency Response (EFR) [30]. For different types of tampering, we show that the EFR is consistent and can therefore be employed to identify



(a) 3×3 average filtering (b) down-sampling by 2 (c) JPEG QF=60 (d) 3×3 median filtering

Figure 6.1: Typical EFRs for four different manipulations. The EFR is shown in a log scale with the center part representing the low-frequency region.

manipulation type.

The EFR of a system $H_X(\boldsymbol{\omega})$ is defined as the ratio of the Discrete-Space Fourier Transform (DSFT) of the system output $Y(\boldsymbol{\omega})$ and the DSFT of the input $X(\boldsymbol{\omega})$, *i.e.* [16],

$$H_X(\boldsymbol{\omega}) = \frac{Y(\boldsymbol{\omega})}{X(\boldsymbol{\omega})}. \tag{6.1}$$

The EFR is input-dependent for non-LSI systems, and when the system is LSI, it coincides with the frequency response. Fig. 6.1 illustrates typical EFRs for different manipulations including (i) down-sampling by 2 (denoted by ↓ 2; the notation ↑ is similarly for up-sampling); (ii) JPEG compression with quality factor (QF) 60, and (iii) 3×3 median filtering (a popular non-linear filter). We obtain similar or "consistent" EFRs for a majority of images in our database; this suggests that even though the EFRs are signal dependent for non-LSI systems, the differences are often minor and similar manipulations produce similar EFRs. In the following, we analyze the reasons behind this consistency for operations such as resampling, JPEG compression, and median filtering.

6.2.1 The EFR for Resampling Operations

Natural images, especially those captured by cameras, possess some implicit structure that may be modified by resampling. Consider an image signal $x(n_1, n_2)$ whose DSFT is denoted by $X(\omega_1, \omega_2)$. The Color Filter Array (CFA) is adopted by most digital cameras for scene sampling. The CFA consists of array of color sensors, each of which captures a corresponding color of the real-world scene at an appropriate pixel location. After sampling, only one color is recorded at each pixel location, and interpolation is performed to recover the lost color information. The Bayer pattern and its variants are commonly used to determine the color to be sensed at each location. Here we consider the green channel for illustration and assume that the green colors are sampled if the following indicator function $p(n_1, n_2)$ is 1:

$$p(n_1, n_2) = \begin{cases} 1 & \text{if } (n_1 + n_2) \text{ is even,} \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

Let $r(n_1, n_2)$ represent the interpolation filter, then the relation between the obtained image signal and the original scene can be expressed as

$$x(n_1, n_2) = [s(n_1, n_2)p(n_1, n_2)] * r(n_1, n_2), \tag{6.3}$$

and in the DSFT domain,

$$X(\omega_1, \omega_2) = [S(\omega_1, \omega_2) * P(\omega_1, \omega_2)]\, R(\omega_1, \omega_2), \tag{6.4}$$

in which $P(\omega_1, \omega_2)$ consists of two impulses at $(0, 0)$ and (π, π), respectively, in a $2\pi \times 2\pi$ period. $P(\omega_1, \omega_2)$ creates a high-frequency image at (π, π) which ideally is eliminated by $R(\omega_1, \omega_2)$. Note that the low-frequency gain of the interpolation filter is $4/2 = 2$ since in a 2×2 grid two pixels will be interpolated from the other two, that is, $R(\omega_1, \omega_2) = 2$ when ω_1 and ω_2 are small. It is well-known [73] that the input-output relation of down-sampling by 2 both in the horizontal and vertical directions in the DSFT domain is given by

$$Y(\omega_1, \omega_2) = \frac{1}{4}\left[X\left(\frac{\omega_1}{2}, \frac{\omega_2}{2}\right) + X\left(\frac{\omega_1 - 2\pi}{2}, \frac{\omega_2}{2}\right) + X\left(\frac{\omega_1}{2}, \frac{\omega_2 - 2\pi}{2}\right) + X\left(\frac{\omega_1 - 2\pi}{2}, \frac{\omega_2 - 2\pi}{2}\right)\right]. \tag{6.5}$$

Assume that the first term dominates the rest in the region $0 \leq \omega_1, \omega_2 \leq \pi$ (that is, when alias can be ignored) and put in (6.4), then the EFR can be expressed as

$$H_X(\omega_1, \omega_2) = \frac{Y(\omega_1, \omega_2)}{X(\omega_1, \omega_2)} \approx \frac{1}{4}\frac{\left[S(\frac{\omega_1}{2}, \frac{\omega_2}{2}) * P(\frac{\omega_1}{2}, \frac{\omega_2}{2})\right] R(\frac{\omega_1}{2}, \frac{\omega_2}{2})}{\left[S(\omega_1, \omega_2) * P(\omega_1, \omega_2)\right] R(\omega_1, \omega_2)}, \quad 0 \leq \omega_1, \omega_2 \leq \pi. \tag{6.6}$$

We model the DSFT of a natural image using the power law decaying, which is suggested by, for example, [71], and claims that the spectrum has a shape in the form of

$$S(\omega_1, \omega_2) = \frac{A}{(\omega_1^2 + \omega_2^2)^{\frac{\alpha}{2}}}, \tag{6.7}$$

for some image-dependent constants A and $\alpha \approx 1$. For low frequencies, *i.e.*, when ω_1 and ω_2 are small, $S(\omega_1, \omega_2) * P(\omega_1, \omega_2) \approx S(\omega_1, \omega_2)$, $S(\frac{\omega_1}{2}, \frac{\omega_2}{2}) * P(\frac{\omega_1}{2}, \frac{\omega_2}{2}) \approx S(\frac{\omega_1}{2}, \frac{\omega_2}{2})$, and $R(\omega_1, \omega_2) \approx R(\frac{\omega_1}{2}, \frac{\omega_2}{2}) \approx 2$, we have

$$H_X(\omega_1, \omega_2) \approx \frac{1}{4}\frac{\left(\frac{A}{((\frac{\omega_1}{2})^2 + (\frac{\omega_1}{2})^2)^{\frac{\alpha}{2}}}\right) \times 2}{\left(\frac{A}{(\omega_1^2 + \omega_2^2)^{\frac{\alpha}{2}}}\right) \times 2} = \left(\frac{1}{2}\right)^{2-\alpha} \approx \frac{1}{2}, \tag{6.8}$$

when $\alpha \approx 1$. That is, the EFR for low frequencies is approximately constant. When either ω_1 or ω_2 goes near π, $S(\omega_1, \omega_2) * P(\omega_1, \omega_2) \approx S(\omega_1, \omega_2)$, $S(\frac{\omega_1}{2}, \frac{\omega_2}{2}) * P(\frac{\omega_1}{2}, \frac{\omega_2}{2}) \approx$

$S(\frac{\omega_1}{2}, \frac{\omega_2}{2})$, and $R(\frac{\omega_1}{2}, \frac{\omega_2}{2}) \gg R(\omega_1, \omega_2)$, so $H_X(\omega_1, \omega_2)$ will be dominated by the ratio of $R(\frac{\omega_1}{2}, \frac{\omega_2}{2})$ and $R(\omega_1, \omega_2)$ and will be large. This is also valid when $R(\omega_1, \omega_2)$ significantly eliminates the high-frequency image of $S(\omega_1, \omega_2)$ near (π, π). In general, however, the behavior of the EFR for high frequencies of ω_1 and ω_2 depends more on the choice of $R(\omega_1, \omega_2)$ and is determined by the camera. Overall, the EFR of down-sampling by 2 will have consistently low values near the low-frequency region, and higher values around high frequencies, as can be observed in Fig. 6.1(b).

Resampling by a general L/M factor can also be analyzed in a similar manner. In this case, we can decompose the resampling operation into the cascade of an up-sampler, $\uparrow L$, a low-pass filter $F(\omega_1, \omega_2)$, and a down-sampler, $\downarrow M$. Note again that the filter $F(\omega_1, \omega_2)$ behaves both as an interpolation filter and a decimation filter and has a low-frequency gain of L^2. Assuming that the aliasing can be ignored, we can derive the approximate EFR as

$$H_X(\omega_1, \omega_2) \approx \frac{F\left(\frac{\omega_1}{M}, \frac{\omega_2}{M}\right)}{M^2} \frac{\left[S(\frac{L\omega_1}{M}, \frac{L\omega_2}{M}) * P(\frac{L\omega_1}{M}, \frac{L\omega_2}{M})\right] R(\frac{L\omega_1}{M}, \frac{L\omega_2}{M})}{\left[S(\omega_1, \omega_2) * P(\omega_1, \omega_2)\right] R(\omega_1, \omega_2)}, \quad 0 \leq \omega_1, \omega_2 \leq \pi. \tag{6.9}$$

At low frequencies, we have

$$H_X(\omega_1, \omega_2) \approx \frac{L^2 \left(\frac{A}{((\frac{L\omega_1}{M})^2 + (\frac{L\omega_1}{M})^2)^{\frac{\alpha}{2}}} \right) \times 2}{M^2 \left(\frac{A}{(\omega_1^2 + \omega_2^2)^{\frac{\alpha}{2}}} \right) \times 2} = \left(\frac{L}{M} \right)^{2-\alpha} \approx \frac{L}{M}. \tag{6.10}$$

And at higher frequencies, the camera-dependent function $R(\omega_1, \omega_2)$ and the resampling-dependent function $F(\omega_1, \omega_2)$ will determine the characteristics of the EFR. Just as in the case of down-sampling by 2, the variations introduced by various cameras do not mask the characteristics of the resampling operation, and thus it is possible to identify the operation exploiting the EFR.

6.2.2 The EFR for JPEG Compression

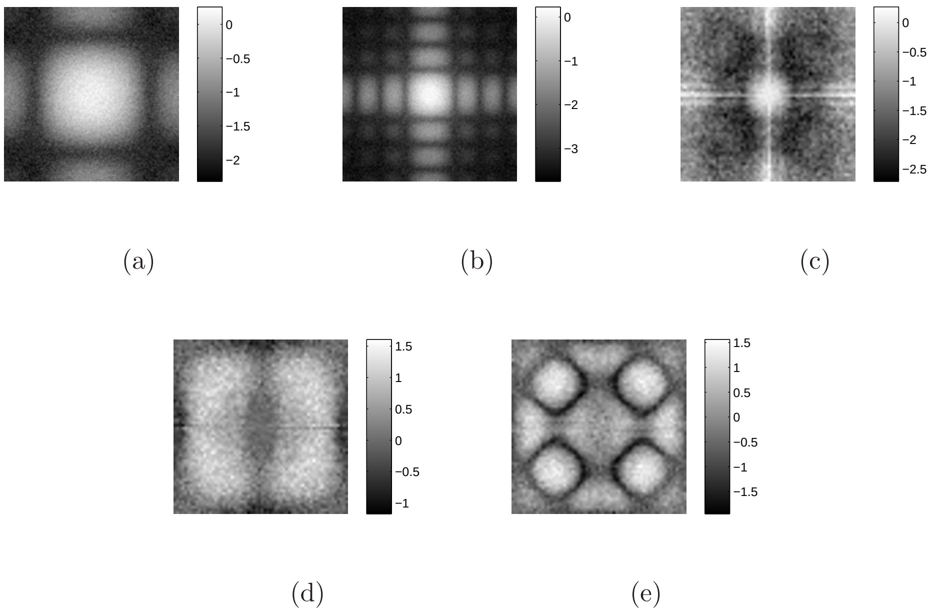
When an image is compressed by JPEG, it is first partitioned into a fixed number of blocks (usually 8×8 or 16×16 pixels), and the Discrete Cosine Transform (DCT) is performed over each block. Each DCT coefficient is quantized into after being divided by its corresponding entry in the quantization matrix then rounded to the nearest integer value. The sequence of quantized DCT coefficients is rearranged in the zig-zag order and losslessly compressed. Decompression is carried out in a reverse order and yields the decompressed image block.

The quantization introduces spectral artifacts that can be manifested by the EFR. First, JPEG compression tends to preserve the low-frequency components by using smaller quantization steps for low-frequency coefficients, which results in smaller quantization error at low frequencies in the DSFT domain. Since low-frequency signal coefficients usually have larger magnitudes, the quantization error will be ignorable compared to the signal magnitude at low frequencies, suggesting that $X(\omega_1, \omega_2) \approx Y(\omega_1, \omega_2)$ and thus $H_X(\omega_1, \omega_2) \approx 1$. For high frequencies, large quantization steps have the effect of destroying image details, that is, $Y(\omega_1, \omega_2) \approx 0$, and thus $H_X(\omega_1, \omega_2) \approx 0$. However, we notice that for certain high frequencies especially those along vertical and horizontal directions, JPEG may increase the resulting coefficient magnitude and thus $H_X(\omega_1, \omega_2) > 1$. This occurs when the quantization error is too large to be ignored but still moderately independent of the signal coefficient. It can also be partially attributed to the rounding error when JPEG performs conversion between floating numbers and integers.

Combining these factors, the EFR of JPEG compression is expected to have values close to unity (or 0 in the log scale) in the low-low frequency region, smaller values in high-high frequency bands, and larger values in the low-high and high-low bands, as is observed in Fig. 6.1(c).

6.2.3 The EFR for Median Filtering

We provide experimental observations about the EFR of median filtering, for three representative cases of input images. First, if the input image has a flat spectrum (*i.e.*, if the input image is white-noise-like), a very strong resemblance between the EFRs of median filtering and average filtering, that is, the sinc-like structure, can be observed, as illustrated in Fig. 6.2(a) and 6.2(b). If a natural image that obeys that power law decaying is used as the input, the central low-frequency parts of the EFR essentially remain, but the mid-frequency and high-frequency regions exhibit some different patterns, as shown by Fig. 6.1(d) and 6.2(c). The resemblance between average filtering and median filtering for frequencies lower than $2\pi/\alpha$, where α is the filter order, has been reported in [30]. Outside this region, more high-frequency coefficients are retained to preserve the signal sharpness. Lastly, if the input image is smooth (*i.e.*, the spectrum only has small high-frequency coefficients), the EFR have large magnitudes for certain mid-frequency coefficients, but its resemblance to that of average filtering is not easily noticeable. We have to remark that, also these EFR patterns are highly consistent in our experimental observations, the theoretical understanding of such consistency still remains to be



(a) (b) (c)

(d) (e)

Figure 6.2: (a) 3 × 3 median filtering with white-noise input; (b) 7 × 7 median filtering with white-noise input; (c) 7 × 7 median filtering with natural image as input; (d) 3 × 3 median filtering with smooth image as input; (e) 7 × 7 median filtering with smooth image as input.

established.

In the next section, we build upon our observation on the EFR consistency across different tampering operations and present a framework for determining the type of tampering operations.

6.3 Tampering Operation Analysis Using EFR

6.3.1 Experiment Setup

In this section, we study the performance of EFR in characterizing different types of tampering operations. As demonstrated in Section 6.2, the EFR is a func-

Table 6.1: Cameras used in our experiment

Brand	Model	Resolution
Canon	PowerShot G7	3648×2736
Canon	PowerShot SD950 IS	4000×3000
Sony	CyberShot DSC-W80	3072×2304
Sony	CyberShot DSC V1	2048×1536
Casio	Exilim EX-Z9	3264×2448
Fujifim	FinePix E550	2848×2136
Olympus	C-5060 WZ	2048×1536

tion of the tampering operation, the camera used, and to some extent dependent on the nature of the input image for non-LSI systems. For example, in the case of resampling, color interpolation coefficients and the low-pass filter are usually a function of the camera and may vary among different camera models. In order to take the effect of the camera into consideration, we employ a data set containing 7 cameras as listed in Table 6.1 in our experiments with 80 images from each camera.

We consider 16 types of manipulations listed in Table 6.2, including resampling, LSI filtering, non-LSI filtering, JPEG compression, and a representative point operation outside these categories. These 16 operations are also grouped into 6 empirical categories consistent with signal-processing knowledge.

Table 6.2: Tampering operations considered in our experiment

Category of Tampering Operation	Configuration
(C1) down-sampling	(O1) $\downarrow 2$
	(O2) $\downarrow 4$
	(O3) $\downarrow 2$ by MATLAB's built-in function imresize
	(O6-O7) 3×3 and 5×5 average filtering followed by $\downarrow 2$
(C2) up-sampling	(O4) $\uparrow 2$ by imresize (bicubic interpolation)
	(O5) $\uparrow 1.5$ by imresize (bicubic interpolation)
(C3) spatial filtering	(O8-O11) 3×3, 5×5, 7×7, and 9×9 average filtering
(C4) non-linear filtering	(O12-O13) 3×3 and 7×7 median filtering
(C5) compression	(O14-O15) JPEG QF=60 and 80
(C6) point-wise operation	(O16) histogram equalization

6.3.2 Robust EFR Estimation for Operation Characterization

We compute EFRs by extracting two corresponding 256×256 blocks from the input and output images, and use the fixed-sized discrete Fourier transform (DFT) to approximate the DSFT. To further reduce the variation within EFRs of the same operation, we perform inverse filtering and Singular Value Decomposition (SVD). Specifically, inverse filtering ensures that all spectral coefficients smaller than a threshold value θ are replaced by θ to avoid numerical instability, and the EFR is decomposed and reconstructed by means of SVD with only N_0 singular value components used. In our setting, $\theta = 0.01$ and $N_0 = 20$.

For resampling operations that change the image sizes, we apply appropriate zero padding in spatial domain to interpolate the frequency components. We preprocess the EFRs by average filtering to reduce the effects of noise, and reduce their dimensionality by first donwsampling it to size 64×64 and then by applying Principal Component Analysis (PCA) to produce 8 features per image.

6.3.3 Identifying the Tampering Type Exploiting EFR Consistency

Fig. 6.3 plots the 2-D principal-component projections of the EFRs for different tampering operations. We notice from Fig. 6.3(a) that operations such as $\downarrow 2$ and 3×3 average filtering exhibit strong inner-operation consistency with the features forming very tight clusters.

The effect of cameras can be studied by capturing the same content using different cameras and examining the consistency of the EFR for different tampering operations. Fig. 6.3(b) shows the 2-D projections of EFR subsets from two cameras of two post-camera manipulations, namely, 7×7 median filtering and JPEG compression with quality factor 80. We see from the figure that the features form four small clusters, but those which belong to the same operation are much closer. This is another level of EFR consistency but still justifies our choice of employing EFRs for tampering type identification.

Lastly, the fact that the EFR partially depends on the input image content as shown in Sec. 6.2.3 suggests a third level of EFR consistency. Fig. 6.3(c) justifies our observation in Sec. 6.2.3 that EFRs corresponding to natural images and smooth images form two clusters close to each other. As long as representative training data from each cluster are available, the distribution of the EFR can be well estimated and thus the EFR can still be employed to identify the manipulation type.

We now examine the classification performance using the EFR as features. We employ Gaussian Mixture Model (GMM) to learn each category and classify the EFR based features using a Maximum-Likelihood (ML) approach. We use the same image set that contains images from 7 cameras; each camera generates 80 images. The 16 operations listed in Table 6.2 are performed on each image. Six image blocks of size 256×256 pixels are then extracted randomly from each of the processed image, and the total number of tampered image blocks is 53760. Note that, in reality, the operations may be performed image-wise but only portions within the image are available for tampering analysis.

We randomly sample 40 training images from each camera for training the classifier and use the remaining images for testing, repeating this process for twenty times to obtain the average performance. Table 6.3 shows the classification performance using the 2-component ML-GMM approach in the form of a confusion matrix. The average classification accuracy is 95.3% suggesting that the EFR can efficiently discriminate between different types of tampering operations.

Table 6.3: Confusion matrix with the original EFR.

%	C1	C2	C3	C4	C5	C6
C1	99.9	0.1	0.0	0.0	0.0	0.0
C2	0.9	99.1	0.0	0.0	0.0	0.0
C3	0.0	0.0	91.2	7.4	1.2	0.2
C4	0.0	0.0	11.8	84.4	3.4	0.4
C5	0.0	0.0	2.3	3.0	94.3	0.4
C6	0.0	0.0	0.0	1.0	3.2	95.8

6.4 Camera Model Identification Using EFR

As discussed in Sec. 6.2.1, the image signal obtained by a camera is partly determined by the exact configuration of the CFA sampling and the interpolation filter. As a result, the EFR of the down-sampling by 2 operation has a dependency on the camera model since such a configuration is camera model specific, especially around high-frequency locations. It is therefore implied that the EFR of $\downarrow 2$ may be used to distinguish betweeen different camera models.

We use the same set of images taken by 7 camera models to verify our hypothesis. In our setting, the entire image will be used to calculate the EFR, which will be further used as the feature for camera model classification. Since calculating the Fourier transform over the entire image is highly computation-consuming, we divide images into blocks of 512×512 pixels. Each block is down-sampled by 2 and we denote the DFT of the ith block before and after $\downarrow 2$ by $X_i(\omega_1, \omega_2)$ and $Y_i(\omega_1, \omega_2)$, respectively. Writing $Y_i(\omega_1, \omega_2) = H(\omega_1, \omega_2)X_i(\omega_1, \omega_2) + E_i(\omega_1, \omega_2)$ where $H(\omega_1, \omega_2)$

stands for the ensemble average EFR of ↓ 2 and $E_i(\omega_1, \omega_2)$ is the modeling inaccuracy, it can be seen that the maximum-likehood estimate of $\|H(\omega_1, \omega_2)\|$ based on N image blocks is given by

$$\|H(\omega_1, \omega_2)\|_{\mathrm{ML}} = \frac{\sum_{i=1}^{N} \|X_i(\omega_1, \omega_2)\| \|Y_i(\omega_1, \omega_2)\|}{\sum_{i=1}^{N} \|X_i(\omega_1, \omega_2)\|^2}, \tag{6.11}$$

if the modeling inaccuracy $E_i(\omega_1, \omega_2)$ is modeled as Gaussian with equal variance across different i. Empirically, we find this estimate more reliable for camera model identification than simple averaging over $\|Y_i(\omega_1, \omega_2) X_i(\omega_1, \omega_2)^{-1}\|$. The EFRs from three color channels are stacked into a feature vector. We randomly select 70 images from each cameras fortraining, and apply PCA to reduce the feature dimension to 32. The obtained 32-dimensional feature vector is linearly scaled so that the feature value falls into $[0, 1]$, and then fed into a 7-class support vector machine (SVM) with a radial basis function kernel. This process is repeated for 500 times.

When the entire EFR is used for camera model classification, the average accuracy of the 7-class classification is 97.40%, which suggests that the EFR of ↓ 2 can effectively distinguish between different camera models. We also consider using low-frequency and high-frequency regions of the EFR for camera model classification. Specifically, two masks $m_L(\omega_1, \omega_2)$ and $m_H(\omega_1, \omega_2)$ are imposed upon the EFR to select the regions that will be used to form the raw feature vector. The two masks are given by $m_L(\omega_1, \omega_2) = \mathbb{1}\{\sqrt{\omega_1^2 + \omega_2^2} \leq R_L\}$ and $m_H(\omega_1, \omega_2) = \mathbb{1}\{\sqrt{\omega_1^2 + \omega_2^2} \geq R_H\}$, respectively, where $\mathbb{1}\{\cdot\}$ is the indicator function, and R_L and R_H are selected so that the two masks have equal sizes of support. With the low-frequency and high-frequency regions selected exclusively for camera model classification, the obtained accuracies are 78.19% and 92.26%, respectively. This is consistent with our observation in Sec. 6.2.1 that the high-frequency region of the EFR carries more camera-dependent information.

6.5 Estimating EFR using Blind Deconvolution

In most applications involving tampering detection, the camera output (namely, the system input) is not accessible and therefore the EFR of the system cannot be readily determined. Nevertheless, blind deconvolution has been shown to be an effective means for estimating the frequency response of a system with only the output signal [69], and in this work, we employ blind deconvoltuion to estimate the EFR.

In principle, any blind deconvolution algorithms that lead to reasonable estimation results can be utilized [37]. We adopt here the one based on the Gaussian prior described in [38]. Specifically, the DSFT of a natural image is modeled as follows:

$$\log p(X(\omega_1, \omega_2)) = -\sum_{\omega_1, \omega_2} \frac{1}{\sigma_X^2(\omega_1, \omega_2)} \|X(\omega_1, \omega_2)\|^2 + C, \tag{6.12}$$

for some normalization constant C. This model is not as accurate as certain alternatives such as the Laplacian prior [38], but is favored here because it can lead to closed-form blind deconvolution solutions. The output of a system modeled as LSI with frequency response $H(\omega_1, \omega_2)$ is given by

$$Y(\omega_1, \omega_2) = H(\omega_1, \omega_2)X(\omega_1, \omega_2) + N(\omega_1, \omega_2), \tag{6.13}$$

where $N(\omega_1, \omega_2)$ stands for the inaccuracy of LSI system modeling and is modeled as Gaussian with zero mean and variance η^2. It can be derived that the output has the following distribution:

$$\log p(Y(\omega_1,\omega_2)) = -\sum_{\omega_1,\omega_2} \frac{1}{\|H(\omega_1,\omega_2)\|^2\sigma_X^2(\omega_1,\omega_2)+\eta^2}\|Y(\omega_1,\omega_2)\|^2 + C', \quad (6.14)$$

for another constant C'. The maximum a posteriori probability estimate of $\|H(\omega_1,\omega_2)\|$ can be found as

$$\|H(\omega_1,\omega_2)\|^2 = \max\left(0, \frac{\|Y(\omega_1,\omega_2)\|^2-\eta^2}{\sigma_X^2(\omega_1,\omega_2)}\right). \quad (6.15)$$

Reference [38] has justified that blind deconvolution based on the Gaussian prior provides a reasonable solution, and we use it here to estimate the EFR. We learn $\sigma_X^2(\omega_1,\omega_2)$ using a diverse set of 1200 images.

6.5.1 Classification using Estimated EFR and Block Fusion

We compare the classification performances of the original EFR and the estimated EFR. Table 6.4 shows the confusion matrix for the estimated EFR using two-component ML-GMM. We notice from the table that the classification accuracy with the estimated EFR is lower for certain manipulation categories compared with the corresponding results obtained with the original EFR reported in Table 6.3. Nevertheless, different categories can still be differentiated effectively using the estimated EFR with an accuracy close to 88.1%, suggesting that the estimation is effective. Further, the fact that a reasonably high accuracy is retained is an indicator that the traces for identifying the tampering are well preserved in the output image, an observation in line with today's blind image forensic research.

As an image is divided into blocks for forensic analysis, some of the blocks may be largely smooth. Such block may have its power concentrated on low-frequency spectral coefficients, and as a result, the EFR estimate for the high-frequency region may appear more noisy. Besides using the robust technique such as SVD described in Sec. 6.3.2 to alleviate such estimation noise, another option is to remove in advance these smooth blocks. We do so by calculating the average Sobel gradient magnitude of each manipulated block, and only keep those blocks with gradient magnitude larger than a threshold g_T which we choose as 10 here. We find that smooth blocks can be effectively identified even if the Sobel operator is conducted after the image is considerbly smoothened. The corresponding confusion matrix is given in Table 6.5, and we can see that the per-category accuracy is increased after smooth blocks are filtered out, and the average accuracy increases by about 3% to 91.0%.

As discussed above, the EFR depends both on the camera and the image content. Such dependence is not desired since it lowers the inner-operation consistency. In this part, we introduce multi-block fusion as a possible means to alleviate such

Table 6.4: Confusion matrix with the estimated EFR.

%	C1	C2	C3	C4	C5	C6
C1	93.5	0.2	1.0	1.5	1.9	1.9
C2	0.6	94.1	0.5	2.9	1.9	0.1
C3	0.4	0.3	95.3	3.7	0.2	0.1
C4	1.9	2.3	8.6	85.2	1.2	0.8
C5	6.4	2.7	0.8	1.0	83.9	5.3
C6	11.0	0.3	0.9	1.7	9.5	76.6

Table 6.5: Confusion matrix with the estimated EFR. Smooth image blocks are removed.

%	C1	C2	C3	C4	C5	C6
C1	93.3	0.1	0.4	1.2	2.8	2.2
C2	0.3	96.8	0.0	1.2	1.7	0.0
C3	0.4	0.1	99.0	0.5	0.1	0.0
C4	1.6	1.4	1.4	93.8	0.9	0.9
C5	6.0	1.2	0.5	1.2	86.1	4.9
C6	11.5	0.0	0.0	1.1	10.1	77.2

dependency. Assuming that the entire image or a certain significant portion of it undergoes the same operation, we can fuse evidence from more than one block to jointly determine the manipulation type.

We adopt the naïve Bayes classifier which assumes that each block of the total N blocks is independent, and the a posteriori probability can be written as

$$P(C_i|F_1, \ldots, F_N) \propto \prod_{j=1}^{N} P(F_j|C_i), \tag{6.16}$$

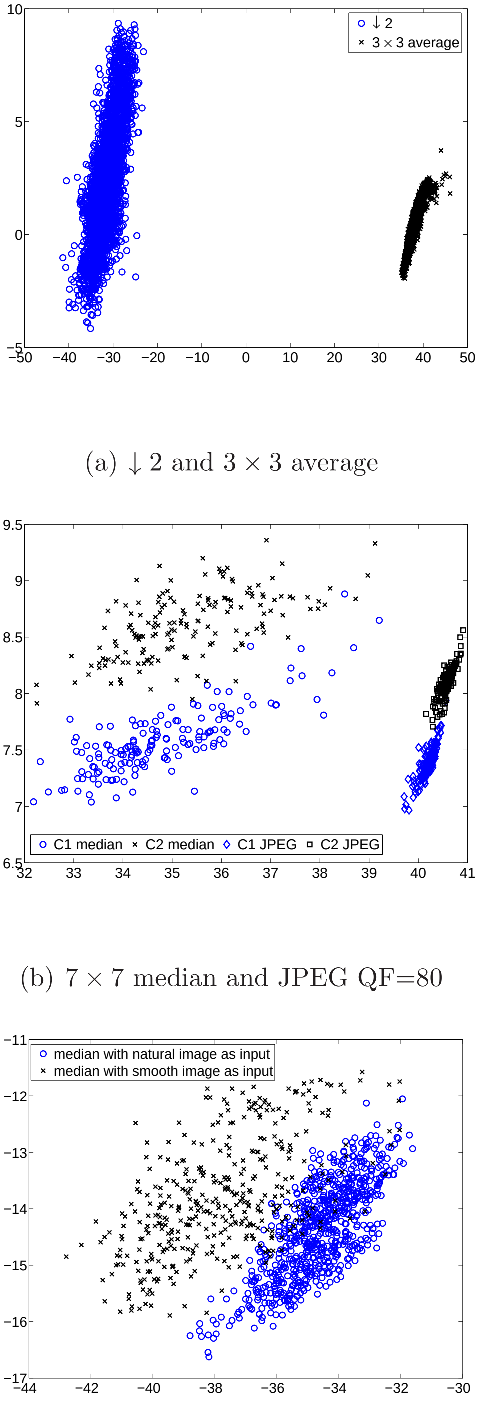
where C_i is the ith category, F_j is the estimated EFR of the jth block. Using the two-component GMM to model $P(F_j|C_i)$, the confusion matrix is shown in Table 6.6. Multi-block fusion improves the average classification accuracy to 97.7%.

6.6 Chapter Summary

In this chapter, we introduce the Empirical Frequency Response (EFR) as a highly universal descriptor for digital tampering operations. We find that many LSI and non-LSI operations exhibit consistencies in the EFR, and therefore the EFR can be utilized to identify tampering operations when the input and output of the tampering module are known. Our results indicate that the proposed EFR based features can classify six categories of tampering with an accuracy of 95.3%. In scenarios where the system input is not available, we show that the EFR can still be estimated just based on the output data, and used for tampering identification with an accuracy of 88.0%; which can be further improved to 97.7% by removing smooth blocks and fusing multiple blocks. Experimental results supported by theoretical reasoning demonstrate the effectiveness of the proposed approach. Besides tampering analysis, we also demonstrate that different properties of the EFR can be used for other forensic applications, such as camera model identification.

Table 6.6: Confusion matrix with the estimated EFR. Non-smooth image blocks in the same image are used jointly to estimate the EFR.

%	C1	C2	C3	C4	C5	C6
C1	98.4	0.0	0.0	0.0	0.7	0.9
C2	0.5	99.2	0.0	0.2	0.2	0.0
C3	0.2	0.0	99.8	0.0	0.0	0.0
C4	0.4	0.0	0.2	99.2	0.1	0.0
C5	0.9	0.0	0.0	0.1	96.4	2.6
C6	3.0	0.0	0.0	0.0	4.1	92.9



(a) ↓ 2 and 3 × 3 average

(b) 7 × 7 median and JPEG QF=80

(c) 7 × 7 median with two types of content

Figure 6.3: Plots showing the 2-D projection of the EFR for (a) down-sampling and average filtering, (b) median filtering and JPEG compression across two different cameras (denoted by C1 and C2), (c) median filtering with two types of content.

CHAPTER 7

Conclusions and Future Perspectives

In this dissertation, we have considered the resiliency of intrinsic fingerprinting and addressed important resiliency issues that arise in real-world intrinsic fingerprinting systems. As we have seen in this dissertation, unfavorable conditions are common, and many current intrinsic fingerprinting that have been designed without awareness of resiliency can easily fail under these conditions or suffer from serious performance degradation. To the best of our knowledge, this dissertation is among the first works that call for explicit efforts to deal with resiliency issues of intrinsic fingerprinting. Upon assessing in depth the vulnerabilities in today's systems, we investigate promising algorithmic solutions to enhancing the resiliency of intrinsic fingerprinting.

Resilience against Content Dependency and Post-processing: First, we consider the content dependency of camera model identification based on color interpolation identification. Our statistical study has shown that the prior schemes exhibit a substantial dependency on the image content. Such dependency can limit the achievable identification performance, which can be further penalized if there exists a mismatch between the training and testing image content. Our study develops profiles that can be used to quantitatively represent the image content, and we propose static and adaptive schemes for selecting training images that fit the content of the testing image. Our experimental results show that by incorporate the notion of content-awareness, the proposed schemes outperform blind schemes and therefore effectively mitigate content dependency and improve the forensic performance.

An interesting research dimension that we plan to investigate next is how to reduce the complexity of image data collection. In our current setting, we assume that a super set of training images is collected for each camera model that is to be identified. In fact, in the entire training process, the collection of training data may demand more time than the execution of the subsequent camera model identification process, and it is of practical interest if one can reduce the time for data collection such as for field use outside full-capacity forensic laboratories. We have begun to consider possible low-complexity training scenarios, including training using printed images and LCD-displayed images. For both cases, we use a digital single-lens reflex camera Canon Rebel T1i to capture a super set of sample images. When training using printed images, we print the sample images at a high resolution (1200 dpi) and use the two cell-phone cameras, a Nokia 6650d model and an Apple iPhone 4 model to retake the training images. The primary difference between the two cell-phone cameras is that the iPhone 4 has an improved short-distance capture capability. As we set the operating false identification rate that a non-targeted camer model is misidentified as the targeted model at 0%, the identification rate associated with Nokia 6650d is 5.8%, whereas the detection rate associated with Apple iPhone 4 reaches 84.2%. These preliminary results suggest that if cameras with short-distance capture capability are available, then printed images can be employed to provide a reasonable identification accuracy with a reduced data collection complexity, otherwise the forensic performance could be severely limited. When training using LCD-displayed images, the sample images are displayed on a 19" LCD screen and the two cell-phone camera models are used in the same way. With the same false identification rate, however, the true identification rates for Nokia 6650d and Apple iPhone 4 are 11.7% and 0%, respectively. An issue with training using displayed images, which has also been found in recent literature [9], is that visible textural patterns can be introduced in recaptured LCD images in the display and recapturing process. It remains challenging how to resolve the effects of such patterns that may bias the color interpolation coefficient estimation.

We also plan to investigate more aspects of the proposed content-aware methodology. For example, are there better ways to integrate the two proposed profiles to improve their representation power? In addition to our proposed profiles, are there other intermediate representations of the image content that can fit a particular identification task? Ultimately, as content dependency is a common issue that can occur in other identification techniques, we believe that the content awareness will have a broader impact, and more forensic tasks can benefit from this design methodology.

The second resiliency issue considered in this dissertation is the post-processing that can take place posterior to the processing module to be identified. When the post-processing becomes strong, it has to be dealt with explicitly. Specifically, we consider in this dissertation the camera identification using imaging noise extracted from low-bit-rate videos. We show that while the compression can significantly distort the extracted noise and reduce the identification accuracy, our proposed frame reordering and unequal weighting can nonetheless leverage the difference in the compression level between different frame types to improve the identification with fewer frames used.

So far, we have mainly considered a hard separation of I and P frames, and assigned weights to each respectively. The optimal strategy for frame reliability estimation will be an interesting direction to be explored. One immediate possibility is to relate the reliability to the PSNR of each video frame, since PSNR carries information about the compression level. The capability of no-reference PSNR estimation [7] will be necessary. It is also beneficial to examine further the underlying reasons that cause the difference in frame PSNR as well as the identification reliability.

Another intriguing direction is to explore the relation between PRNUs associated with different resolutions. In particular, most of today's digital cameras support both image and video outputs, using the same array of sensor. Since videos usually have a smaller resolution than images, some kind of down-sizing of the whole sensor output values must be used. It has been observed that the aspect ratios of the image and video outputs are usually different, which implies that extra cropping is also involved. We are interested in understanding the internal resizing mechanism of a camera when only a completely or semi non-intrusive approach can be employed. The knowledge of the internal resizing as well as cropping may be helpful in developing effective forensic techniques for resized signals. For example, as digital images often undergo less severe compression than videos on the same camera, the PRNU pattern extracted from images may be more reliable. So if the mapping between images and videos associated with the same camera can be established, then we anticipate that a high-quality PRNU pattern extracted from images can be remapped to identify video frames and achieve a higher accuracy.

Resilience against Anti-Forensics: As there are always adversaries who have the incentives to counteract forensic investigations, we then study another resiliency aspect of intrinsic fingerprinting: how well it can resist anti-forensics. Two practical intrinsic fingerprinting schemes are considered for this study. First, we propose anti-forensic techniques, referred to as parameter perturbation and algorithm mixing, which aim at circumventing and misleading forensic identification of color interpolation algorithms, respectively. Both techniques can be applied to a wide range of color interpolation algorithms. Parameter perturbation is particularly powerful when interpolation involves direction classification, which is very common in advanced interpolation algorithms. Our investigation of algorithm mixing, on the other hand, shows that linearly mixing interpolated pixels from two independent interpolation algorithms can not only reduce the identification performance but also preserve the image quality. Our study sheds light on the inherent vulnerabilities of current color interpolation identification systems, which further motivate countermeasures as well as adjustments of anti-forensics. We characterize such an interaction as a color interpolation identification game, and derive the optimal strategies for forensic analysts and adversaries using game-theoretic techniques.

We also study the anti-forensics resiliency of a recent forensic technique based on electrical network frequency (ENF) analysis that can determine the creation time of digital recordings. This technique detects the frequency fluctuations of the ENF signal and compares the fluctuations to the references measured at the power network to determine the creation time and region. Our work is the first that examines plausible anti-forensic operations that can manipulate the ENF signal and the corresponding space-time information. We also establish a mathematical framework for ENF anti-forensics to understand its detectability, which motivates detection schemes such as inter-frequency consistency and spectrogram consistency checks as well as improved anti-forensics via envelope adjustment based on Hilbert Transform. We characterize the dynamic interplay between forensic analysts and adversaries using an evolutionary perspective and a game-theoretic perspective, and provide quantitative studies of representative scenarios and derive the optimal strategies.

As far as we know, our study of ENF anti-forensics is the first work in this research direction. In light of the potential deployment of the ENF-based forensic techniques, we envision that as more upcoming anti-forensics and countermeasures are to appear, our framework can help understand and analyze these advancements.

A fundamentally important goal of ENF and anti-ENF research is to obtain an improved understanding of the formation mechanism of the ENF signal. Such an understanding will facilitate the characterization of current techniques, and lead to more sophisticated forensic and anti-forensic methods.

Anti-forensics in itself is a subject that has much more to explore. On one hand, it is important to examine the performance of more forensic techniques under adversarial settings, and seek feasible improvements that can address identified vulnerabilities. On the other hand, studies of individual techniques may lead to a holistic understanding of anti-forensics and countermeasures, and we anticipate that more comprehensive theoretical frameworks can be established as well. Readers who are interested in other case studies and relevant discussions of anti-forensics can refer to [35, 61].

Universal Identification of Intrinsic Fingerprints: Our study of the empirical frequency response (EFR) provides a new way to address another resiliency aspect of intrinsic fingerprinting: to identify operations in a (nearly) universal manner. So far, we have considered linear shift-invariant (LSI) operations as well some other non-LSI operations, such as resampling, compression, non-linear filtering, and even some point-wise operations. While different types of operations may have different levels of consistency in terms of their EFR representations, we find that the EFR is highly discriminative in separating different operation categories, and also promising for operations that are within the same category but have different parameters.

An interesting research issue that remains to be better understood is on whether more operations exhibit EFR consistency, under what conditions and to what extent. In practice, multiple operations may form a chain, and it is still an open question in a general setting how the EFR corresponding to the entire processing chain is related to EFRs of individual operations. As we employ blind deconvolution to estimate the EFR when the input image is unavailable, how the EFR estimation depends on the accuracy of blind deconvolution is also an interesting research issue. Exploring such issues will lead to a deeper understanding of the EFR. Finally, the recent notion of forensic hash [42] aims at attaching compact side information to digital images to assist forensic tasks beyond tampering detection. Since the EFR offers the capability of characterizing LSI and certain non-LSI operations, it has the potential to serve as a representation component that can be combined into the framework of forensic hash to support finer identification of processing history.

Systematic Framework: As this dissertation studies important issues that must be considered in the real-world use of intrinsic fingerprinting, continued effort should be devoted to a systematic framework for the fingerprint resiliency. Toward this end, proper definitions for resiliency that can be measured quantitatively can be established, accompanied by a resiliency evaluation methodology that examines comprehensively how a given fingerprinting scheme performs under various testing conditions and in the presence of adversaries. A goal of resilient intrinsic fingerprinting is to devise fingerprint extraction and matching methods that are invariant or insensitive to deviations of operational conditions. One example concerns visual signals whose derived fingerprints may be different if the signals' spatial resolution is changed. Since most of todays digital cameras support outputs of multiple resolutions and resizing is a common content-preserving operation, it is highly desirable if intrinsic traces extracted at different resolutions could be properly matched to one another. The resiliency framework can benefit from the theoretical framework of component forensics [70], which has modeled different identification settings of digital devices and derived corresponding identification performances. We expect that in a signal processing chain where a component's effects may be masked by subsequent components, the identification resiliency of the particular component can be evaluated using existing results from component forensics. Finally, it is of interest to see if anti-forensics can be incorporated into the framework.

Bibliography

[1] Wikipidia article on“Mains hum”.

[2] J. Adams. Interaction between color plane interpolation and other image processing functions in electronic photography. In *SPIE Cam. and Sys. for Elec. Photo. & Scien.c Imag.*, Feb. 1995.

[3] J. Adams and J. Hamilton. Adaptive color plane interpolation in single sensor color electronic camera (US Patent #5,506,619), 1996.

[4] I. Avcibas, S. Bayram, N. Memon, M. Ramkumar, and B. Sankur. A Classifier Design for Detecting Image Manipulations. In *Proc. of Int. Conf. on Image Process. (ICIP)*, volume 4, pages 2645–2648, Oct. 2004.

[5] S. Bayram, H. T. Sencar, N. Memon, and I. Avcibas. Source camera identification based on cfa interpolation. In *Proc. of International Conf. Image Proc.*, 2005.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[7] T. Brandao and M. P. Queluz. No-reference PSNR estimation algorithm for H.264 encoded video sequences. In *Proc. of 16th Euro. Sig. Proc. Conf.*, Lausanne, Switzerland, Aug. 2008.

[8] A. Buades, B. Coll, J. M. Morel, and C. Sbert. Self-similarity driven color demosaicking. *IEEE Trans. on Image Processing*, 18(6):1192–1202, June 2009.

[9] H. Cao and A. Kot. Identification of recaptured photographs on LCD screens. In *Proc. of IEEE Inter. Conf. on Acoustics, Speech, and Signal Proc.*, March 2010.

[10] H. Cao and A. C. Kot. Accurate detection of demosaicing regularity for digital image forensics. *IEEE Trans. on Information Forensics and Security*, 4(4):899 –910, Dec. 2009.

[11] H. Cao and A. C. Kot. Mobile camera identification using demosaicing features. In *Proc. of IEEE International Symposium on Circuits and Systems*, 2010.

[12] M. Chen, J. Fridrich, M. Goljan, and J. Lukas. Source digital camcorder identification using sensor photo-response non-uniformity. In *Proc. of SPIE Electronic Imaging*, pages 1G–1H, Photonics West, Jan. 2007.

[13] X. Chu, M .C. Stamm, W. S. Lin, and K. J. R. Liu. Forensic identification of compressively sensed images. In *Proc. of Int. Conf. Acoustic, Speech, and Signal Processing*, 2012.

[14] W.-H. Chuang, R. Garg, and M. Wu. How secure are power network signature based time stamps? In *Proc. of ACM Conference on Computer and Communications Security*, 2012.

[15] W.-H. Chuang, H. Su, and M. Wu. Exploring compression effects for improved source camera identification using strongly compressed video. In *Proc. of IEEE International Conference on Image Processing*, pages 1953 –1956, Sep. 2011.

[16] W.-H. Chuang, A. Swaminathan, and M. Wu. Tampering identification using empirical frequency response. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1517–1520, April 2009.

[17] W. H. Chuang and M. Wu. Semi non-intrusive training for cell-phone camera model linkage. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2010.

[18] W.-H. Chuang and M. Wu. Robustness of color interpolation identification against anti-forensic operations. In *Proc. of 14th Information Hiding Conference*, 2012.

[19] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, 3 edition, 2002.

[20] H. Farid. Blind Inverse Gamma Correction. *IEEE Trans. on Image Process.*, 10(10):1428–1433, Oct. 2001.

[21] H. Farid. Image forgery detection. *IEEE Signal Proc. Magazine*, 26(2):16 –25, March 2009.

[22] H. Farid. Seeing is not believing. *IEEE Spectrum*, 8(46):44–48, 2009.

[23] H. Farid and S. Lyu. Higher-Order Wavelet Statistics and Their Application to Digital Forensics. In *IEEE Workshop on Statistical Analysis in Computer Vision*, June 2003.

[24] J. Fridrich. Digital image forensics. *IEEE Signal Proc. Magazine*, 26(2):26 –37, March 2009.

[25] R. Garg, A. L. Varna, and M. Wu. "Seeing" ENF: Natural time stamp for digital video via optical sensing and signal processing. In *Proc. of ACM Multimedia*, Nov. 2011.

[26] H. Gou, A. Swaminathan, and M. Wu. Intrinsic sensor noise features for forensic analysis on scanners and scanned images. *IEEE Trans on Infor. Forensics and Security*, 4(3):476 –491, Sep. 2009.

[27] C. Grigoras. Applications of ENF criterion in forensics: audio, video, computer, and telecommunication analysis. *Forensic Science International*, 167:136–145, Apr. 2007.

[28] III H. Daumé and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Int. Res.*, 26:101–126, May 2006.

[29] S. Haykin. *Communication Systems*. Wiley Publishing, 5th edition, 2009.

[30] T. S. Huang, editor. *Two-Dimensional Diginal Signal Processing II: Transforms and Median Filters*. Springer-Verlag, 1981.

[31] M. K. Johnson and H. Farid. Exposing digital forgeries in complex lighting environments. *IEEE Transactions on Infor. Forensics and Security*, 2(3):450 –461, Sep. 2007.

[32] E. Kee and H. Farid. Exposing digital forgeries from 3-d lighting environments. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2010.

[33] N. Khanna, A. K. Mikkilineni, and E. J. Delp. Scanner identification using feature-based processing and analysis. *IEEE Trans. on Infor. Forensics and Security*, 4(1):123 –139, Mar. 2009.

[34] M. Kirchner and R. Bohme. Synthesis of color filter array pattern in digital images. In *Proc. of SPIE-IS&T Electronic Imaging: Media Forensics and Security*, 2009.

[35] M. Kirchner and R. Bohme. Counter-forensics: Attacking image forensics. In H. T. Sencar and N. Memon, editors, *Digital Image Forensics*. Springer, 2012.

[36] R. Klinkenberg and T. Joachims. Detecting concept drift with support vector machines. In *In Proc. of the Seventeenth Int. Conf. on Machine Learning*, pages 487–494, 2000.

[37] D. Kundur and D. Hatzionakos. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64, May 1996.

[38] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[39] C. T. Li. Source camera identification using enhanced sensor pattern noise. *IEEE Trans. on Infor. Forensics and Security*, 5(2):280–287, June 2010.

[40] X. Li, B. Gunturk, and L. Zhang. Image demosaicing: A systematic survey. In *Visual Communications and Image Processing, Proc. of the SPIE*, volume 6822, 2008.

[41] Y. Long and Y. Huang. Image based source camera identification using demosaicing. In *Proc. of MSP*, 2006.

[42] W. Lu, A. L. Varna, and M. Wu. Forensic hash for multimedia information. In *Proc. of SPIE Media Forensics and Security*, 2010.

[43] J. Lukas and J. Fridrich. Estimation of Primary Quantization Matrix in Double Compressed JPEG Images. In *Proc. of the Digital Forensics Research Workshop*, Aug. 2003.

[44] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Trans. on Infor. Security and Forensics*, 1:205–214, Jan. 2006.

[45] W. Luo, M. Wu, and J. Huang. Mpeg recompression detection based on block artifacts. In *Proc. of SPIE Conf. on Security, Forensics, Steganography, and Watermarking of Multimedia Contents*, volume 6819, January 2008.

[46] S. McCloskey. Confidence weighting for sensor fingerprinting. In *Proc. of IEEE Conf. on Com. Vision and Patt. Rec. Workshops*, 2008.

[47] C. McKay, A. Swaminathan, H. Gou, and M. Wu. Image acquisition forensics: Forensic analysis to identify imaging source. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2008.

[48] R. Mislan. Cellphone crime solvers. *IEEE Spectrum*, 47(7):34–39, Jul. 2010.

[49] Roger B. Myerson. *Game Theory: Analysis of Conflict.* Harvard University Press, 1991.

[50] C. Nachenberg. Computer virus-antivirus coevolution. *Communications on the ACM*, 40:46–51, Jan. 1997.

[51] R. Ng. *Digital Light Field Photography.* PhD thesis, Stanford University, Jul. 2006.

[52] T. T. Ng, S. F. Chang, J. Hsu, L. Xie, and M. P. Tsui. Physics-motivated features for distinguishing photographic images and computer graphics. In *Proc. ACM Multimedia*, 2005.

[53] T. T. Ng, S. F. Chang, Y. F. Hsu, L. Xie, and M. P. Tsui. Physics-motivated features for distinguishing photographic images and computer graphics. In *Proc. of ACM Multimedia*, Singapore, November 2005.

[54] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time Signal Processing*. Prentice-Hall, 2 edition, 1999.

[55] D. Paliy, V. Katkovnik, R. Bilcu, S. Alenius, and K. Egiazarian. Spatially adaptive color filter array interpolation for noiseless and noisy data. *International Journal of Imaging Systems and Technology*, 17:503–513, 2007.

[56] D. Parikh and K. Grauman. Relative attributes. In *Proc. of the International Conference on Computer Vision (ICCV)*, 2011.

[57] A. C. Popescu and H. Farid. Exposing digital forgeries in color filter array interpolated images. *IEEE Trans. on Signal Proc.*, 53(10):3948–3959, October 2005.

[58] A.C. Popescu and H. Farid. Exposing Digital Forgeries by Detecting Traces of Re-sampling. *IEEE Trans. on Signal. Process.*, 53(2):758–767, Feb. 2005.

[59] D. P. N. Rodriguez, J. A. Apolinario, and L. W. P. Biscainho. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Trans. on Information Forensics and Security*, 5(3):534 –543, Sep. 2010.

[60] R. W. Sanders. Digital authenticity using the electrical network frequency. In *Proc. of 33rd AES Int. Conf. on Audio Forensics, Theory and Practice*, Jun. 2008.

[61] M. C. Stamm. *Digital Multimedia Forensics and Anti-Forensics.* PhD thesis, University of Maryland, College Park, 2012.

[62] M. C. Stamm, W. S. Lin, and K. J. R. Liu. Forensics v.s. anti-forensics: A decision and game-theoretic framework. In *Proc. of International Conf. on Acoustics, Speech, and Signal Processing*, Mar. 2012.

[63] M. C. Stamm, W. S. Lin, and K. J. R. Liu. Temporal forensics and anti-forensics for motion compensated video. *IEEE Trans. on Information Forensics and Security*, 7(4):1315 –1329, Aug. 2012.

[64] M. C. Stamm and K. J. R. Liu. Anti-forensics of digital image compression. *IEEE Transactions on Infor. Forensics and Security*, 6(3):1050 –1065, Sep. 2011.

[65] M. C. Stamm and K. J. R. Liu. Anti-forensics of digital image compression. *IEEE Transactions on Information Forensics and Security*, 2011.

[66] Y. Sun and Y. Liu. Security of online reputation systems: The evolution of attacks and defenses. *IEEE Signal Processing Magazine*, 29(2):87 –97, Mar. 2012.

[67] A. Swaminathan, M. Wu, and K. J. R. Liu. Nonintrusive component forensics of visual sensors using output images. *IEEE Trans. on Infor. Forensics and Security*, 2(2):91–106, March 2007.

[68] A. Swaminathan, M. Wu, and K. J. R. Liu. Digital image forensics via intrinsic fingerprints. *IEEE Trans. on Infor. Forensics and Security*, 3(1):101–117, March 2008.

[69] A. Swaminathan, M. Wu, and K. J. R. Liu. Forensic Analysis via Intrinsic Fingerprints. *IEEE Trans. on Infor. Forensics and Security*, 3(1):101–107, March 2008.

[70] A. Swaminathan, M. Wu, and K. J. R. Liu. Component forensics. *IEEE Signal Proc. Magazine*, 26(2):38–48, March 2009.

[71] A. Torralba and A. Oliva. Statistics of Natural Image Categories. *Network: computation in neural systems*, 14:391–412, 2003.

[72] A. Tsymbal. The problem of concept drift: Definitions and related work. Technical report, Department of Computer Science, Trinity College: Dublin, Ireland, 2004.

[73] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[74] W. Wang and H. Farid. Exposing digital forgeries in video by detecting double mpeg compression. In *Proc. of ACM Multimedia and Security Workshop*, Geneva, Switzerland, 2006.

[75] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13(4):600 –612, Apr. 2004.

[76] Wikipedia. Condition number — Wikipedia, the free encyclopedia.

[77] T. F. Wu, C. J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2003.

[78] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proc. of ACM Multimedia*, 2007.

[79] J. Yu, S. Craver, and E. Li. Toward the identification of DSLR lenses by chromatic aberration. In *Proc. of SPIE Conference on Media Forensics and Security*, 2011.

[80] L. Zhang, X. Wu, A. Buades, and X. Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Elec. Imag.*, (023016), Apr. to Jun. 2011.